Physical, Transcriptional and Comparative Mapping on the Human X Chromosome

Thesis

Physical, Transcriptional and Comparative Mapping on the Human X Chromosome

by

Gareth Rhys Howell

Thesis submitted for the

degree of Doctor of Philosophy

The Open University

1st March 2002

The Wellcome Trust Sanger Institute

Wellcome Trust Genome Campus

Hinxton

Cambridge, UK

DATE OF SUBMISSION: 7 MARCH 2002
DATE OF AWARD: 19 JUNE 2002

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

This thesis is dedicated to my wife, Andrea

and to my daughters, Megan and Cadi.

Abstract

Progress in the study of the human genome has resulted in the production of a complete clone map and associated 'working draft' sequence. This will underpin the completion of the sequence itself, the annotation of genes and other features, and application of this new found knowledge. This thesis focuses on the evolving methods to determine the map, and to use the emerging sequence for the study of genes, incorporating new studies of other genomes to enhance progress in understanding and interpretation to biology and medicine. The success of the endeavours is necessarily accompanied by the development and evolution of new technologies and by critical assessment of the progress in acquiring knowledge of the genomic information.

Evolution of mapping technologies included the development of the larger insert bacterial cloning systems (PACs) and (BACs), and an increase in available landmarks both from YAC maps and RH maps. The work described in chapter 3, followed this evolution and was applied to construct a 6 Mb sequence-ready bacterial clone contig map in Xq22. A minimum set of clones was chosen for genomic sequencing. The resulting sequence map was compared to previously published maps and analysed both for common repeats, and previously unidentified low copy repeats.

The availability of the emerging sequence of the human genome provided a resource for identification of the features encoded within. In chapter 4, the sequence of a 7 Mb region in Xq23-24 was analysed for the presence of genes using a combination of sequence similarity searches against both protein and DNA databases, and *ab initio* gene prediction. Predicted genes were confirmed where possible, by generating novel

cDNA sequence. The region contained 33 confirmed genes (of which 14 were confirmed during this study), 11 predicted genes and 20 pseudogenes.

Comparative genome sequence analysis is a powerful method both for aiding human gene identification and identifying other features encoded within the human genome such as regulatory elements. Comparing the genomes of two or more species also provides insights into the evolution of the species since the divergence from a common ancestor. Sequence from a 1 Mb region in human Xq24 was compared in two other species, mouse and zebrafish. In chapter 5, bacterial clone contigs for sequencing were constructed in the mouse by designing mouse-specific STSs orthologous to human sequence for clone isolation. In chapter 6, bacterial clone contigs for sequencing were constructed in zebrafish using STS from exons of human genes to identify zebrafish BAC clones by reduced stringency hybridisation. Comparative analysis of the region showed that humans and mice are more highly conserved than humans and zebrafish, in terms of gene content and organisation. A combination of comparative sequence analysis tools identified 14 novel potential conserved sequences between human and mouse, one of which was also conserved in zebrafish.

Acknowledgements

I would like to thank my supervisors David Bentley and Mark Ross for all the help, advice and guidance provided throughout the course of this thesis. On a number of occasions career development 'chats' were required to ensure it kept on track and for that I am most grateful. You have both ensured I have stayed focussed whilst doing a part-time PhD.

A number of people have made invaluable contributions, without which this thesis would not have been possible. Particular thanks go to The Sanger Institute sequencing teams: Darren Graffham, Christine Bird and other members of the X chromosome finishing team, particularly because Xq22 proved a horrid region to finish. Christine spent many hours toying with cosmids, solving finishing problems caused by duplications and deletions – it wasn't my fault! Adrienne Hunt also deserves a special mention for work on both human, but also zebrafish clones, rushing finishing through to provide me with some sequence to work with. All FISH analyses were carried out by the FISH group and in particular, Pawandeep Dhami, whose life would have been much easier if I had not required 'just a few' fibrefish experiments! Jackie Bye for cDNA resources, general advice and proof reading. Sarah Hunt and Carol Scott for informatics support. The Chromosome 22 group, including John Collins, Dave Beare and Ian Dunham - "Much of the sequence analysis in this thesis would not have been possible without the advice and pearl scripts from these three wise men" (is that what you told me to put?).

Thanks go to the members of the Experimental Gene Annotation Group; Graeme, Jackie, Kevin, Liz and Kate, for allowing me the time to finish this thesis. You have all kept the group going. Maybe you'll see me in the lab soon?

Personal thanks go to my friends. Pod, I promised you a special mention and hopefully this is good enough. You have taught me so much over the years about all sorts of things and supported me all the way. Your skills relating to lab work and thesis writing, reading and checking were invaluable. You certainly know how to proof read a thesis! Also, your unprintable one-liners have kept the amusement levels high! Ian and Tamsin, fellow PhD students, for your energies, discussions and proof reading and also for cheering me up when on the odd occasion I have felt a little bit grumpy. Apologies to Ian for deleting his gene structures and changing the Xq22 map at will. Cords, without the tea and bacon sandwich interruptions, thesis writing would have been much duller. Dave, you were there at the start of the thesis (1400 cosmids – thanks!) and at the end (proof reading), thanks for all your help. And Simon – you have been there as a friend and confidant throughout my time at the Sanger Institute, without your straight talking and competitiveness who knows where I would be?!

Final thanks go to my family: My wife Andrea, who has been there for me throughout and believed in me all the way. You have given encouragement whenever needed, even when not always understanding a word of what I was saying! And Dad and Mum, for your will, passion and enthusiasm.

To every one mentioned, and no doubt many I've omitted, I owe you a pint (or a half – I am Welsh after all!)

Table of Contents page

List of Figures:

Chapter 4

Chapter 5:

List of Tables:

Glossary of Abbreviations

ACeDB	*A C. elegans* database
ANT2	adenosine nucleotide transporter 2
Alu-PCR	*Alu*-element-mediated polymerase chain reaction
ATP (dATP, ddATP)	adenosine 5'-triphosphate (deoxy-, dideoxy-)
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
BLIXEM	BLAST In an X-windows Embedded Multiple Alignment
β-ME	β-mercaptoethanol
bp	base pair
BSA	bovine serum albumin
BTK	Bruton's tyrosine kinase
°C	degrees Celsius
cDNA	complementary deoxyribonucleic acid
chr	chromosome
cM	centiMorgan
cm	centimetre
CpG	cytidyl phosphoguanosine dinucleotide
cpm	counts per minute
cR	centiRays
CTP (dCTP, ddCTP)	cytidine 5'-triphosphate (deoxy-, dideoxy-)
dbEST	database of expressed sequence tags
DNA	deoxyribonucleic acid
dNTP	2'-deoxyribonucleoside 5'-triphosphate
DTT	dithiothreitol
EDTA	ethylenediamine tetra-acetic acid
EMBL	European Molecular Biology Laboratory
EST	expressed sequence tag
FISH	fluorescence *in situ* hybridisation
FP	forward primer
FPC	Fingerprinting Contig

g	gram
G banding	Geimsa banding
GDB	Genome Database
GSC	Genome Sequencing Centre, St Louis
GTP (dGTP, ddGTP)	guanine 5'-triphosphate (deoxy-, dideoxy-)
HGMP	Human Genome Mapping Resource Centre
HGP	Human Genome Project
HMM	Hidden Markov Model
HPRT	hypoxanthine phosphoribosyltransferase
kb	kilobase pairs
l	litre
LAMP2	lysosomal-associated membrane protein 2
LB	Luria-Bertani
LD	linkage disequilibrium
LINE	long interspersed nuclear element
M	molar
Mb	megabase pairs
µg	microgram
µl	microlitre
µM	micromolar
min(s)	minute(s)
mg	milligram
ml	millilitre
mm	millimetre
mM	millimolar
MRX	X-linked non-specific mental retardation
NSMR	Non-specific mental retardation
NCBI	National Centre for Biotechnology Information
ng	nanogram
nm	nanometre
O/N	overnight
OD	optical density
OMIM	On-line Mendelian Inheritance in Man

PAC	P1-derived artificial chromosome
PAR	pseudoautosomal region
PCR	polymerase chain reaction
PFAM	Protein Family
PFGE	pulsed-field gel electrophoresis
pg	picogram
plp	proteolipid protein
PMD	Pelizaeus Merchbacher Disease
pmol	picomole
poly(dT)	poly-deoxyribothymidyl oligonucleotide
R banding	Reverse Geimsa banding
RH	radiation hybrid
RFLP	restriction fragment length polymorphism
RNA (mRNA, rRNA, tRNA)	ribonucleic acid (messenger-, ribosomal-, transfer-)
RP	reverse primer
Rnase A	ribonuclease A
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcription polymerase chain reaction
SDS	sodium dodecyl sulphate
sec(s)	second(s)
seq	sequence
SINE	short interspersed nuclear element
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
SSPCR	single-sided specificity PCR
STS	sequence tagged site
TEMED	N,N,N',N'-tetramethylethylenediamine
TrEMBL	Translated EMBL
TIGR	The Institute of Genome Research
Tris	tris(hydroxylmethyl)aminomethane
U	unit
UTR	untranslated region
uv	ultraviolet

V	volt
v/v	volume/volume
VNTR	variable number of tandem repeats
W	watt
w/v	weight/volume
Wash U.	Washington University
WGS	whole genome shotgun
Xace	X chromosome version of ACeDB
XCI	X chromosome inactivation
XIC	X-inactivation centre
Xist	X inactive specific transcript
XLA	X-linked agammaglobulinaemia
XLMR	X-linked mental retardation
XLP	X-linked lymphoproliferative disease
YAC	yeast artificial chromosome

Publications

Parts of this work presented in this thesis have appeared previously in the following publications:

Bentley, D. R., Deloukas, P., Dunham, A., French, L., Gregory, S. G., Humphray, S. J., *et al.* (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409:** 942-3.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Chapter 1

Introduction

Introduction

The evolution of mapping and sequencing strategies and methodologies is enabling the complete sequence of many genomes to be elucidated. The methods that were used to generate clone maps and sequence of the genomes of smaller model organisms have provided the foundation for analysing the larger and more complex genomes of vertebrates. The DNA sequence of many species, from simple viruses to more complex multicellular organisms such as the nematode worm and the fruitfly, is now available and the finishing of the human genome sequence is scheduled for 2003. The generation of the complete DNA sequence of the mouse and the zebrafish genomes is now well underway. The availability of the genomic sequence facilitates the identification of the biologically significant units encoded within, such as genes and regulatory elements. The genome sequence of different organisms enables comparisons to be made between them in order to both improve the identification of the functional units and to aid in the understanding of their biological significance.

1.1 Mapping and sequencing of model organisms

The ability to generate comprehensive maps and very accurate DNA sequence of large genomes, for example, the human genome, has been made possible because of the pioneering work carried out on smaller genomes. Some of the key organisms for which DNA sequence is now available are listed in Table 1.1.

Table 1.1: *Complete DNA sequence*

Organism	Size	Method	Reference
Bacteriophage ϕX174	5 kb	Plus and Minus	Sanger, F., *et al.*, (1977a) Sanger, F., *et al.*, (1978)
Bacteriophage λ	48 kb	Random seq	Sanger, F., *et al.*, (1982)
H. influenzae	1830 kb	Whole Genome Shotgun	Fleischmann, R. D., *et al.*, (1995)
E. coli (K12)	4,600 kb	Clone-based	Blattner, F. R., *et al.*, (1997)
S. cerevisiae	12,000 kb	Clone-based	Goffeau, A., *et al.*, (1996)
C. elegans	97,000 kb	Clone –based	*TCSC (1998)
D. melanogaster	120,000 kb	Whole Genome Shotgun	Adams, M. D., *et al.*, (2000)
A. thaliana	125, 000 kb	Clone-based	**TAGI (2000)

* TCSC – C elegans Sequencing Consortium, The
** TAGI – Arabidopsis Genome Initiative, The

The first genome to be sequenced was that of bacteriophage ϕX174 using the plus-minus method based on the elongation of DNA chains with DNA polymerase. However this method was laborious and prone to errors (Sanger, F., *et al.*, 1977a), and so was completely sequenced using chain terminators (Sanger, F., *et al.*, 1977b, Sanger, F., *et al.*, 1978) . Bacteriophage λ was the first organism to be sequenced using the strategy based on sequencing random pieces of DNA (the 'shotgun' approach) (Anderson, S. 1981), in this case restriction fragments (Sanger, F., *et al.*, 1982).The first bacterial genome to be completely sequenced, *Haemophilus influenzae* (*H. influenzae*) (Fleischmann, R. D., *et al.*, 1995), was sequenced using the whole genome shotgun strategy developed for the bacteriophage λ project. In the case of *H. influenzae*, random small insert (2 kb) and large insert (15-20 kb) libraries were constructed and was followed by high throughput DNA sequencing, assembly, sequence editing and annotation.

For the analysis of more complex genomes, a strategy involving the initial generation of a physical map was developed. The discovery of site-specific restriction endonucleases led to the widespread use of restriction mapping for understanding the organisation of DNA fragments. For regions spanning less than 50 kb, restriction maps were constructed routinely. In a few cases, maps as large as 600 kb were constructed, but it proved difficult to extend existing mapping methods beyond that range (Olson, M. V., *et al.*, 1986). For mapping larger regions such as the genomes of *Caenorhabditis elegans* (*C. elegans*) (Coulson, A., *et al.*, 1986) and *Saccharomyces cerevisiae* (*S. cerevisiae*) (Olson, M. V., *et al.*, 1986) two strategies using restriction enzymes were implemented.. For the generation of clone maps covering the *S. cerevisiae* genome, a redundant set of lambda clones was analysed using a single restriction digest, and fragment sizes for each clone compared (Olson, M. V., *et al.*, 1986). For the generation of clone maps covering the *C. elegans* genome, a much larger genome than had been attempted previously, a strategy using cosmids (Collins, J., *et al.*, 1978) was developed. Whole genome restriction digest fingerprinting of cosmids picked at random, and representing a six-fold coverage of the *C. elegans* genome, was carried out. Cosmid DNA was digested with *Hin*d III, the ends of the fragments were labelled with a radioactive molecule and a secondary digest of the labelled *Hin*d III fragments with *Sau*3AI generated fragments of a size that could be resolved on a denaturing polyacrylamide gel. Cosmids with similar fingerprints were interpreted as containing overlapping cloned inserts, and assembled into contigs. In total, 17,500 cosmids were assembled into 700 contigs (Coulson, A., *et al.*, 1986; Coulson, A., *et al.*, 1988). The gaps which persisted between contigs after all the cosmids had been analysed were bridged with yeast artificial chromosomes (YACs) (Burke, D. T., *et al.*, 1987; Coulson, A., *et al.*, 1988). The

YACs had the advantage that DNA fragments greater than that possible for cosmids could be cloned (the YACs were approximately 200 kb in size). The YACs that bridged gaps between contigs accounted for approximately 20% of the nematode genome.

The sequencing of the *C. elegans* genome was carried out on a clone by clone basis and in two phases. In the first phase, each cosmid was subcloned into phage vectors (1.3-2 kb insert size) and the resulting subclones sequenced at random. These random sequences were assembled automatically into sequence contigs. At this point, contigs greater than 2 kb were released into the public domain. The second phase involved a targeted finishing process that enabled contiguous pieces of highly accurate DNA sequence representing the original cosmid to be generated. Finishing involved closing gaps between sequence contigs with targeted resequencing, resolving ambiguities such as GC tracts and improving low quality sequence. Fosmids, a bacterial-based cloning system similar to cosmids but maintained in a single copy number (Kim, U. J., *et al.*, 1992), were incorporated and sequenced in regions where there was no cosmid coverage. Finally, YAC clones were sequenced to close gaps between bacterial clone contigs and in 1998 the completion of the DNA sequence of the nematode worm, the first multicellular organism was announced (C. elegans Sequencing Consortium, The, 1998). Other genomes sequenced by the clone map-based approach included *Escherichia coli* (*E. coli*) (Blattner, F. R., *et al.*, 1997; Kohara, Y., *et al.*, 1987), *S. cerevisiae* (Olson, M. V., *et al.*, 1986; Goffeau, A., *et al.*, 1996) and the first plant *Arabidopsis thaliana* (*A. thaliana*) (Arabidopsis Genome Initiative, The, 2000). *Drosophila melanogaster* (*D. melanogaster*), was sequenced

by whole genome shotgun in part as a proof of principal for sequencing the human genome (Adams, M. D., *et al.*, 2000) (discussed later – see Section 1.2).

1.2 Mapping and sequencing the human genome

The advances in the mapping and sequencing of the smaller genomes of organisms such as the nematode worm created the possibility that the complete DNA sequence of the human genome could be generated. The human genome is approximately three gigabases (Gb), which is 30 times larger than that of the worm. It is divided onto twenty-two pairs of autosomal chromosomes and two sex chromosomes (either XY in males or XX in females). One average-sized human chromosome is roughly equivalent to the *C. elegans* genome (100 Mb). The human chromosomes were first characterised by cytogenetic mapping using differential staining and visualisation to identify unique banding patterns for each chromosome. Metaphase chromosome preparations can be treated with trypsin digestion and Giemsa staining and biologically significant differences were observed between the bands (see Table 1.2).

Table 1.2: *Comparison of G-bands and R-bands*

G-bands (Paleogenome)	R-bands (Neogenome)
Dark Staining	Light Staining
Late replicating	Early replicating
Early condensation	Late condensation
DNase insensitive	DNase sensitive
Less frequent recombination	More frequent recombination

There are up to 850 bands that can be visualised by staining (Bickmore, W. A., *et al.*, 1989) and the unique banding enables the identification of each chromosome in the human genome.

At the time the sequencing of the nematode worm began in 1989, small regions of the human genome were being studied because of a specific interest in that region such as association with a particular disease or phenotype. The discovery and characterisation of polymorphic markers within the genome and the development of methods for analysing linkage between them in pedigree studies led to the development and formalisation of genetic mapping. Markers on a genetic map must exist in two or more forms (or alleles) and are said to be polymorphic. This allows for a distinction between different alleles on different chromosomes in a population or different individuals. Naturally occurring DNA polymorphisms are present throughout the human genome and have been utilised for the production of genetic maps.

Genetic maps relate the distance between markers to the likelihood of recombination occuring during meiosis. The closer together two landmarks are on a chromosome, the less likelihood there is of a recombination occuring at meiosis. The reverse is true for markers that are further apart. Distance on a genetic map is measured in centimorgans (cM) and is a measure of the frequency of recombination between two markers. Recombination frequency can be used as an approximate measure of physical distance and is based on the assumption that recombination is random, i.e. there is an equal chance that it will occur at any position in the genome. On this basis, one centimorgan is equivalent to a 1% recombination frequency. The human genome covers 3000 cM and is 3000 Mb in size, therefore 1 cM is corresponds on

average to 1 Mb. However recombination is known to be non-random so this is subject to inaccuracy (Dib, C., *et al.*, 1996).

The first genetic map covering the human genome was based on restriction fragment length polymorphisms (RFLP) (Donis-Keller, H., *et al.*, 1987), but the low frequency of their occurrence, and the maximum heterozygosity of 50% of RFLPs limited their usefulness. The discovery of hypervariable regions in DNA showing mulitallelic variation within the population (Wyman, A. R., *et al.*, 1980) provided a new source of markers with major advantages over RFLPs. Both mini-satellite markers (11-60bp repeats) and micro-satellite markers (di-, tri- and tetra-nuceotide repeats) have been utilised in genetic mapping (Nakamura, Y., *et al.*, 1987), (Weissenbach, J., *et al.*, 1992), although mini-satellites have an irregular distribution, being more prevalent near telomeres and therefore are less suitable for genetic mapping. The first high resolution, genetic map of the human genome was produced in 1992 (Weissenbach, J., *et al.*, 1992) using micro-satellites. Subsequent maps followed in 1994 (Buetow, K. H., *et al.*, 1994; Murray, J. C., *et al.*, 1994; Gyapay, G., *et al.*, 1994) culminating in the version published in 1996 (Dib, C., *et al.*, 1996) contained 5,264 short tandem repeats with a mean heterozygosity of 70%. The sex-averaged distance covered equals 3,699 cM and the average distance between markers is 1.6 cM. However, a subset of 2032 markers is ordered at high odds (1000:1 or better odds against an alternative order), and these maps have allowed localisation of many monogenic disorders.

The ability to order landmarks across the human genome was becoming increasingly important with the growing interest in genome-wide analysis. A method for marker

ordering, similar in some respects to genetic mapping, is hybrid mapping. In the same way that genetic maps rely upon breaks in chromosomes during meiosis, hybrid mapping is based on either naturally occurring physical breaks, such as translocations or deletions, or breaks induced artificially, for example by irradiation.

Naturally occuring chromosomal abnormalities such as translocations, deletions or inversions are useful mapping tools. Somatic cell hybrids have been generated from fusions between human cells containing abnormal chromosomes and rodent cells (e.g hypoxanthine phosphoribosyltransferase (HPRT) negative), where the fusion cell containing the aberrant human chromosome has been recovered using selectable markers (e.g - HPRT, which was used as a selectable marker for incoming human DNA). Panels of somatic cell hybrids have been generated containing different regions of the same chromosome. DNA from each hybrid can be tested for the presence or absence of particular landmarks. Combining the data from different hybrids allows the landmarks to be placed in intervals defined by overlapping chromosome segments, for example in chromosome 22 (Bell, C. J., *et al.*, 1995).

A panel of somatic cell hybrids containing aberrant chromosomes is limited to naturally occurring abnormalities. However techniques are available to induce chromosome breaks randomly. Radiation hybrids (RH) were originally developed by Goss and Harris in 1975, where they fragmented chromosomes by irradiation-induced breakage (Goss, S. J., *et al.*, 1975). Fragments were then rescued by fusion to a rodent cell and hybrids isolated by selection. Multiple cell lines each contain pieces of the human genome, together in overlapping fragments. Individual breaks define order where the closer two markers are together the more likely they are to be

maintained on a single fragment. For the analysis of two markers, a matrix is generated and the separation of markers is measured by the number of differences between the two matrices. Markers can be screened against a panel of these hybrids to determine their position in the genome relative to each other and other markers whose location is known. The greater the amount of radiation used, the greater the number of fragments generated, and subsequently the greater the short-range resolution of the panel. Similarly to genetic mapping, the RH map includes a set of framework markers that are ordered at high odds (greater than 1000:1 against a different order). Other markers have then been positioned in 'bins' relative to these framework markers. For instance, radiation panels derived from the whole human genome, such as the GB4 panel and the G3 panel (Gyapay, G., *et al.*, 1996; Stewart, E. A., *et al.*, 1997) have been used to position more than 30,000 ESTs and other genetic and physical markers to produce a gene map of the human genome (Deloukas, P., *et al.*, 1998, updated electronically 1999, see http://www.ncbi.nih.nlm.gov/genemap99).

The plans to sequence the human genome relied upon the construction of clone-based maps. The YAC system was the largest possible cloning system and provided potentially the most efficient way to tackle the problem of assembling long-range clone maps of the larger human genome. YAC maps could be constructed using the markers that had been positioned on both the available genetic and hybrid maps. YAC libraries of the human genome (Anand, R., *et al.*, 1990; Imai, T., *et al.*, 1990; Larin, Z., *et al.*, 1991; Chumakov, I. M., *et al.*, 1992b; Albertsen, H. M., *et al.*, 1990) were generated and improvements in the cloning systems led to the construction of libraries containing clones with an insert size that exceeded 1 Mb (Chumakov, I. M.,

et al., 1995). The first human YAC contigs were built using shared STSs across disease-associated loci such as the cystic fibrosis gene (Green, E. D., *et al.*, 1990) and the dystrophin gene (Coffey, A. J., *et al.*, 1992) and subsequently larger regions were covered e.g the euchromatic portions of chromosomes Y, 21 and 22 (Foote, S., *et al.*, 1992), (Chumakov, I., *et al.*, 1992b), (Collins, J. E., *et al.*, 1995). The success of these maps lay in the combination of applying unique STSs to map YACs by STS content mapping and in the vectorette end-rescue of YACs (Riley, J., *et al.*, 1990) to isolate new markers from the ends of clones which could be used to develop probes for walking.

However, YACs have two major disadvantages. They are prone to deletions and rearrangements, possibly because of the size of the cloned insert and because of the recombinogenic background of the commonly used yeast hosts. Their other major problem is chimaerism, which is the occurrence of insert DNA from two non-contiguous regions of the human genome. Chimaerism may result either from co-ligation of two fragments to form a single insert, or recombination between two independent recombinants co-transformed in the same host cell (Monaco, A. P., *et al.*, 1994). The STS content approach to mapping YACs substantially overcame these disadvantages for constructing long-range maps. The absence of an STS or a group of STSs in a single YAC indicated the presence of a deletion in the YAC clone. In the case of chimaerism, each STS generated from the end of a YAC was checked by PCR as being from the region of interest where possible, or at least on the same chromosome (by screening a chromosome-specific hybrid cell line).

The lack of a sufficient density of markers to generate YAC maps across the whole genome led Bellane-Chantelot C., *et al* (1992) to develop YAC fingerprinting to compensate for the landmark deficiency (Bellanne-Chantelot, C., *et al.*, 1992). Subsequently, an attempt to generate a YAC-based map covering the human genome using a combination of landmark mapping and YAC fingerprinting was published in 1993 (Cohen, D., *et al.*, 1993). Fingerprints for 33,000 YAC clones were generated by detecting fragments containing medium-repeat sequences and assembled into contigs on the basis of similarly sized fragments. More than 2000 genetic markers, and 5322 novel STSs generated from sequencing PCR products amplified between *Alu* elements, were also used to screen the YACs and detect overlaps between clones. Finally, approximately 500 YACs containing genetically mapped polymorphic STSs (one every 7.4 cM) were positioned on metaphase chromosomes using FISH. The method attempted to mirror the successful approach developed for mapping the nematode worm, using large scale fingerprinting to assemble clones into contigs but with additional organisation from positioning the contigs in the genome with markers, including the polymorphic markers placed on the genetic map. An updated version of the map was published in 1995 (Chumakov, I. M., *et al.*, 1995). However, even though small regions of the YAC map were assembled correctly, the inherent problem of chimaerism assembled non-contiguous portions of the human genome within the same contigs. Combined with the false negative and positive data, this caused major misassemblies of the data, and resulted in a poor final map. Given the inconsistencies that still existed in the YAC maps across the whole genome, construction of YAC maps across single chromosomes continued as groups had more confidence in the data they were generating and resulted in YAC maps covering chromosomes 3 (Gemmill, R. M., *et al.*, 1995), 7 (Bouffard, G. G., *et al.*, 1997), 12

(Krauter, K., *et al.*, 1995), 16 (Doggett, N. A., *et al.*, 1995), 21 (Chumakov, I. M., *et al.*, 1992b), (Nizetic, D., *et al.*, 1994), 22 (Collins, J. E., *et al.*, 1995), X (Nagaraja, R., *et al.*, 1997) and Y (Foote, S., *et al.*, 1992).

A more successful attempt to generate a physical map across the human genome was carried out using a combination of RH mapping and STS-based YAC mapping (Hudson, T. J., *et al.*, 1995). The physical map contained 15,086 STSs, of which 10,850 were screened against YACs to produce an integrated map anchored by RH and genetic maps. This approach to YAC mapping was more successful than the previous attempts by Cohen, D., *et al.*, (1993) as it did not rely on fingerprinting to assemble YACs into contigs.

Although some of the YACs that formed part of the clone map covering the nematode worm were being sequenced, the inherent problems of chimaerism, deletions and the difficulty in purifying the YAC DNA from the host DNA, made YAC clones a less suitable substrate for large-scale sequencing than bacterial clones. The majority of the map covering the nematode genome was constructed using the bacterial clones. Initial sequence-ready bacterial clone contig construction in human utilised some of the techniques developed for the nematode worm. For instance, for the early work on sequencing human chromosome 22, although cosmids were identified using the YACs and STSs ordered on the YAC map as a framework, the cosmids were assembled into contigs using restriction digest fingerprinting (see Section 1.1). The development of larger insert bacterial cloning systems (P1 artificial chromosomes - PACs (Ioannou, P. A., *et al.*, 1994) and bacterial artificial chromosomes - BACs (Shizuya, H., *et al.*, 1992), with insert sizes of between 130-

150 kb for PACs, and up to 300 kb for BACs, improved the efficiency of map construction. Combined with the density of markers available from the YAC maps bacterial clone contigs could be generated directly from the markers covering the whole of chromosome 22. The availability of a high density of STSs from YAC maps also enabled bacterial clone contig construction to proceed on other chromosomes at an earlier stage, notably the X and Y chromosomes, and chromosome 7 and 21.

Sequencing of the bacterial clones in the maps such as those on chromosome 22 also utilised the procedures that had been developed during the sequencing of smaller genomes such as the nematode worm (see Section 1.1). A minimally overlapping set of bacterial clones were chosen for sequencing on a clone by clone basis, and a random shotgun phase was followed by a targeted finishing process for each clone. The complete sequence of chromosome 22, the first human chromosome to be finished, was published in 1999 (Dunham, I., *et al.*, 1999).

The early construction of bacterial clone contigs on chromosomes such as 7, 16, and 22, and the X and Y chromosome relied upon the high density of ordered markers available from the chromosome-specific YAC maps. For the chromosomes for which no YAC map was available, a sufficient density of markers was achieved from a combination of existing STSs, positioned on available genetic and radiation hybrid maps, and in some cases, novel STSs generated from chromosome-specific sequences (Ross, M., *et al.*, 1997). Radiation hybrid mapping of these STSs produced maps with a sufficient density (average 15 markers per Mb) of ordered STSs that could be used as a framework for map construction using PACs or BACs,

and this eliminated the need to generate YAC maps. Chromosome mapping projects that were among the first to benefit from this increased density of markers included chromosomes 1, 2, 6, 19, 20 and 21. The clone map of chromosome 6 is now complete and sequencing is well advanced. The finished sequence of chromosome 20 was recently announced (Deloukas, P., *et al.*, 2001) with the sequence of chromosome 6 expected to be finished in the spring of 2002.

Other human chromosomes were not covered by sequence-ready maps and in order to generate contigs covering the entire human genome, human genome BAC libraries (initially RPCI-11 and RPCI-13) were fingerprinted using a restriction digest fingerprinting method similar to that developed for the nematode worm. The method consisted of a HindIII digest of BAC DNA, with the fragments resolved on an agarose gel (closely resembling the approach taken to map the *S. cerevisiae* genome, (Olson, M. V., *et al.*, 1986). Analysis of the fingerprints was carried out using IMAGE and FPC, as was the case for fluorescent fingerprinting. Incorporation of the mapping data already available enabled minimum tiling sets of bacterial clones to be identified representing more than 90% of the human genome.

All of the human sequence generated by early in 2000 had been carried out using the clone-based strategy. Previous sequencing of whole genomes had centred around two strategies, clone-based or whole genome shotgun (see Section 1.1). However, the increased complexity of the human genome led to a debate as to whether the whole genome shotgun strategy was suitable for sequencing such a genome (Green, P., 1997, Venter, J. C., *et al.*, 1998, Weber, J. L., *et al.*, 1997). The clone-based approach makes full use of the map information to verify each clone to be sequenced

and allows efficient co-ordination to minimise the effort and duplication. For complex genomes, it also avoids major assembly problems due to genome-wide repeats. The whole genome shotgun approach circumvents the construction of clone-based maps and involves generating and assembling random sequence reads from cloned fragments varying in size from 1-2 kb upto 50 kb into sequence contigs or scaffolds, and integration of sequences generated at the ends of BAC clones. In spite of the concerns, a whole genome shotgun strategy was implemented to generate human genomic sequence in a useable form in under two years. Celera, in collaboration with the Berkley *Drosophila* Genome project had used the whole genome shotgun strategy to generate sequence representing approximately 120 Mb of the euchromatic portion of the *D. melanogaster* genome (Adams, M. D., *et al.*, 2000).

Celera (the private effort) was to make the human sequence available in the form of a database which researchers could subscribe to for a fee. The ethos of the Human Genome Project (HGP - the public effort) was to make all the human sequence freely available in the public domain as it was produced. The increasing interest and demand in providing as much sequence of the human genome as quickly as possible in the public domain to support new research led to the formalisation of the existing intermediate, i.e. the assembled shotgun sequence data of each clone, as a defined product, and it was termed the 'working draft sequence'. Draft sequence mostly contains very accurate sequence (an error rate of less than 1 in 10,000 bases), with some gaps and missassemblies. In February of 2001, the completion of the draft sequence, the first stage of sequencing the human genome, was announced (Lander, E. S., *et al.*, 2001) and provided researchers with human sequence covering up to

90% of the genome, two or more years earlier than the anticipated production of finished sequence.

At the same time Celera announced the completion of the whole genome shotgun sequence assembly of the human genome in 100 kb scaffolds (Venter, J. C., *et al.*, 2001). A comparison showed that although there was more raw sequence in the Celera data (99% as opposed to 94%), there were fewer contigs in the HGP data and they contained a higher percentage of assembled sequence. One major advantage of the clone-based approach is the ability to carry out a targeted finishing process to produce a high quality product with as few a gaps as possible, and this effort is continuing. To date, the complete sequence of chromosomes 20 (Deloukas, P., *et al.*, 2001), 21 (Hattori, M., *et al.*, 2000) and 22 (Dunham, I., *et al.*, 1999) has been determined, with other chromosomes such as chromosome 6 and chromosome 14 soon to be announced. The completion of the human genome sequence is planned for spring of 2003.

1.3 Interpreting the human genome sequence

The complete DNA sequence of the human genome contains all the information required for the correct development, structure and function of a human being. The identification of the features embodied in the human genome sequence will be one step towards the understanding of the processes that generate and sustain life. As an essential part of this process, the human genome sequence will provide the basis for the identification of all human genes, their organisation and physical position within

the genome. The activation and suppression of transcription of these genes are controlled by molecules binding to promoter sequences (usually located immediately upstream of genes) and regulatory elements (which can be located within, nearby or far away from the gene or genes they influence).

Current estimates suggest that at least 50 % of the human genome is made up of repeat sequences (analysis carried out on the draft sequence, (IHGSC, 2001)). These can be identified in human sequence by comparing genomic sequence to databases of prototypic sequences representing repetitive DNA from primates (see http://ftp.genome.washington.edu/cgi-bin/RepeatMasker and http://www.girinst.org/Repbase_Update.html).These can be broadly divided into five classes, (1) transposon-derived interspersed repeats, (2) inactive retroposed copies of genes (processed pseudogenes), (3) simple sequence repeats, (4) segmental duplications and (5) tandemly repeated sequences such as those seen at centromeres and telomeres, which are involved in the maintenance of chromosomes. By far the most common is the transposon-derived interspersed repeat which accounts for more than 80% of all repeats currently recognised, and includes short interspersed elements (SINES) and long interspersed elements (LINES). Analysis of the draft sequence showed segmental duplications at the pericentromic and subtelomeric regions that are present elsewhere in the genome.

It is not fully understood what the remainder of the human genome encodes for, but will include features such as origins of replication and may include sequences such as those specific to chromosome packaging, but an important first goal is to identify all the genes encoded with the human genome.

1.3.1 Gene identification

A gene can be defined as a region of genomic DNA that is transcribed to form a functional RNA molecule. Some of these RNA molecules are processed to form messenger RNAs (mRNAs) and translated to form proteins. Others function within the cell as RNA molecules, such as those that are associated with the ribosome (ribosomal RNAs, rRNAs or transfer RNAs, tRNAs), or the spliceosome (e.g snoRNAs). All genes will have a transcription start site, and those mRNAs that code for proteins will also have a translation start site and a translation termination site. Individual genes of simple organisms such as prokaryotes are located within a single stretch of genomic DNA, but genes of the more complex organisms, from yeast onwards, are composed of exons interrupted by introns which are removed from the RNA molecule during processing steps that occur after transcription. The process of gene identification utilises these features involved in the correct functioning of the genes notably codon usage and splice junctions, as they can be identified within the genomic sequence, along with other features associated with genes such as CpG islands.

Estimates for the number of genes in the human genome have varied widely depending on the type and the interpretation of information used. These are summarised in Table 1.3.

Table 1.3: *Genes in the human genome*

Data set	Gene Number	Date	Reference or source
Hypothetical	100,000	1992	Gilbert, W., *et al.*,1992
CpG Islands	80,000	1993	Antequera, F., *et al.*, 1993
EST clusters	60,000-70,000	1994	Fields, C., *et al.*, 1994
Unigene clusters	92,000	1996	Schuler, G. D., *et al.*, 1996
Gene sequences	140,000	1999	*IncyteGenomics
Chr. 22 seq.	43,000-61,000	1999	Dunham, I., *et al.*, 1999
Chrs 22, 21 seq.	44,000	2000	Hattori, M., *et al.*, 2000
Tetraodon seq.	28,000-34,000	2000	Roest-Crollius, H., *et al.*, 2000
ESTs in dbEST	120,000	2000	Liang, F., *et al.*, 2000
EST and mRNA	35,000	2000	Ewing, B., *et al.*, 2000
Draft Sequence	31,000	2001	IHGSC, 2001

*press release available at http://incyte.com/company/news/1999/genes.shtml

Estimates for the number of genes encoded in the human genome have varied depending upon the type of information available to calculate the figure. An early rough estimate suggested a genome size of 3,000 Mb could contain 300,000 non-overlapping units of 10 kb, or 100,000 units of 30 kb (Gilbert, W., 1992). A later estimate of gene number by Antequera and Bird (1993) used the fact that the 5' end of approximately 56% of genes are associated with CpG islands, which are regions of non-methlyated DNA that contain the dinucleotide CG at the expected frequency. Analysis of CpG islands present in small amounts of sequence available at the time, suggested the human genome contains approximately 80,000 genes (Antequera, F., *et al.*, 1993). Fields *et al* (1994) clustered the available EST data being generated as part of an EST sequencing project to suggest that there may be between 60,000 and 70,000 genes in the human genome (Fields, C., *et al.*, 1994). There have been two estimates, based on cDNA sequence (Incyte Genomics) and EST clusters in dbEST that the human gene number exceeds 100,000 (Liang, F., *et al.*, 2000). However, more recent estimates based on larger datasets or more refined analyses tend to converge on a lower figure, in the region of 30,000, as follows. Analysis of the

finished sequence of human chromosome 22, which accounts for 1.1% of the human genome and was predicted to contain a minimum of 679 genes, would suggest there are between 43,000 and 61,000 genes (Dunham, I., *et al.*, 1999). This figure may be artificially inflated by the incorporation of 134 pseudogenes, which assuming they are not expressed, are not likely to be represented in the other estimates. In combination with the analysis of human chromosome 21, the figure is quoted to be approximately 44,000, but this figure assumes chromosomes 21 and 22 represent an average gene density similar to that observed across the genome (Hattori, M., *et al.*, 2000). This is slighty higher than a study carried out which looked at conserved sequences between human and tetraodon and suggested there may be as few as 28,000 genes present (Roest-Crollius, H., *et al.*, 2000). The most recent prediction, based on the analysis of the human draft sequence suggests the human genome may contain 31,000 genes (IHGSC, 2001).

Prior to large-scale sequencing of the human genome, gene identification relied upon the identification of an mRNA representing a particular gene. The ability to generate complementary DNA (cDNA) from an mRNA molecule using reverse transcriptase enabled the cloning of the cDNA molecule for further investigation. For instance, the cloning and partial sequencing of the cDNAs representing the α–, β–, and γ–globin genes enabled characterisation of the structure of the globin gene cluster (Little, P. F., *et al.*, 1979; Rabbitts, T. H., 1976). cDNA libraries containing clones representing the complement of mRNA molecules from individual tissues were also generated (Gubler, U., *et al.*, 1983).

Individual or multiple genes of interest could be sequenced, primarily by using a poly(dT) primer to prime from the polyA tail present at the 3' end of the cDNA clones. The advances in the sequencing technologies including the anchored poly(dT) primer (Khan, A. S., *et al.*, 1991) and the improvements in the cDNA library preparation led to the development of strategies for large scale single-pass sequencing and mapping of human cDNAs. It was suggested the sequencing of cDNA clones would be a more efficient strategy for the identification of all human genes, rather than mapping and sequencing the entire human genome (Brenner, S., 1990). Projects involving the partial sequencing of cDNA clones, primarily representing the 3' end of genes, were generating large amounts of novel expressed sequences, such as that described by Adams, M.D., *et al.* (1991) who generated brain-specific expressed sequence tags (ESTs) representing more than 300 novel genes (Adams, M. D., *et al.*, 1991). In 1992 it was reported that global cDNA sequencing may have identified as many as 10,000 human genes (Khan, A. S., *et al.*, 1992). By 1996, there were approximately 450,000 ESTs residing in Genbank (Hillier, L. D., *et al.*, 1996). These ESTs were clustered into non-redundant sets (Unigene, Boguski, M. S., 1995; Schuler, G. D., *et al.*, 1996), Merck Gene Index (Aaronson, J. S., *et al.*, 1996), TIGR Human cDNA collection (Adams, M. D., *et al.*, 1995). Each cluster includes all sequence information, as well as any expression and mapping data that are available. There are currently 1.2 million ESTs in 84,000 clusters in the Unigene collection (December 2001). Sequence alignment programs are used to identify sequences matching at greater than 97%. This is likely to be an over estimate of the overall gene number as one gene could be represented by more than one EST cluster, particularly for the genes with large mRNAs which may be represented by an EST cluster at both the 5' and 3' end.

The EST sequencing projects which used a wide variety of cDNA libraries and carried out single pass sequencing from both the 5' end and 3' end of cDNA clones, proved extremely useful for gene identification but as a stand alone attempt to identify all the genes in the human genome had major shortcomings. cDNA sequence alone gives no indication of the structure and organisation of a gene such as the positioning and size of introns, nor does it provide sequence of the elements involved in regulation of gene expression. There is also evidence that the cDNA libraries used for EST sequencing may have contained contaminating genomic DNA, from which priming for sequencing could have occurred. Also, some mRNAs contained intronic sequences which were subsequently incorporated in to the EST data (Burglin, T. R., *et al.*, 1992). Single pass sequencing from the 5' end and the 3' end of each cDNA clone may not generate the entire sequence coverage particularly for larger cDNA clones. The generation of the cDNA from the mRNA using reverse transcription was not always 100% efficient and so the cDNA libraries did not necessarily represent the full length mRNA molecule. Sequencing from the 5' end of a truncated cDNA clone can give a false indication of the true 5' end of a gene. If a gene is not expressed in any of the cDNA libraries used for EST sequencing, the gene will never be identified from these datasets.

Human genome sequence enables a whole new set of information that, although previously known, could not be applied, such as sequence motifs, homology searches and computational predictions and provides the basis for a more systematic approach to gene identification. The identification of genes in genomic sequence can be divided into four main categories: (1) direct evidence of transcription based on EST and cDNA sequence, (2) comparative protein analysis as proteins or parts of proteins

look like other genes in both human and non-human sequence, (3) *ab initio* gene prediction, (4) comparing genome sequences of different organisms on the assumptions that regions that have conserved function since the divergence from a common ancestor will remain conserved at the sequence level.

(1) Direct evidence

The comparison of cDNA sequence with the corresponding genomic sequence is a more powerful method of identifying genes and their structures, than cDNA sequence alone which gives no information regarding gene structure and organisation. The availability of the genomic sequence and the ability to align the vast amounts of both full length and partial cDNA sequences enables not only the identification of the genes, but also their exon/intron structure. However, in order to identify all genes by this manner, a cDNA representing each gene would need to be sequenced. The importance of this strategy led to the establishment of a number of full length cDNA sequencing efforts to complement the genomic sequencing efforts, for example Mammalian Gene Project (MGP – http://mgc.nci.nih.gov/) and RIKEN (http://genome.gsc.riken.go.jp/home.html). Unlike single pass cDNA sequencing which provided partial information, these more recent cDNA sequencing projects are working to produce sequence covering the entire length of the cDNA clones.

Full-length cDNA sequencing is still limited by the quality and diversity of the cDNA libraries used. Improvements have been made in the technologies used to generate full-length cDNA clones, including the efficiency of the reverse transcriptase enzymes used and the procedure for selecting full length clones. For instance, the cap-trapper method has been developed based on the introduction of a

biotin group into the diol residue of the cap structure of the mRNA, followed by RNase I treatment to select full-length cDNA clones (Carninci, P., *et al.*, 2001). Clones that are not full length will loose the biotin group and not be trapped using streptavidin-coated magnetic beads. However, the cDNA sequence may still not represent the entire length of the gene, identifying only part of the gene, and if the gene is not expressed, or expressed at a low level, in the cDNA libraries used, the gene will not be identified at all by this approach.

(2) Comparative protein analysis

Genes encode proteins which are made up of discrete domains that can be defined on the basis that they are members of recognisable families with specific structure and/or function, which work in combination to contribute to the overall functioning of the protein. For instance the prothrombin prescusor contains a gla, two kringle domains, and one trypsin domain (Bateman, A., *et al.*, 2002) although in some cases a protein may only contain a single domain, for example *SH2D1A* which contains a single SH2 domain (Coffey, A. J., *et al.*, 1998).

Genes can be clustered into families based on sequence similarity at the nucleotide and amino acid level, but also based on the similarity of the domains they contain. The important genetic event in generating a gene family can be either; (1) The divergence of a common ancestral organism to form two or more species with related genes or orthologues which are free to evolve separately but will remain similar at the sequence level if their functions remain similar, (2) The duplication of an ancestral gene to form paralogues within a species, and possible expansion to form

gene clusters to produce related genes, so they can evolve independently for new biological functions.

Non-homologous recombination between repeated elements is one method by which gene duplication can occur. After gene duplication, alterations in the nucleotide sequence may lead to either an altered function of one gene or a silencing of one gene to form a pseudogene (Papadakis, M. N., *et al.*, 1999). An example of a gene family is the globin genes. The α-globin genes, of which there are three functional copies and two pseudogenes located in a cluster on chromosome 16, are highly similar at the nucleotide sequence level, and are suggested to have arisen from a single ancestral α-globin gene by tandem duplication (Papadakis, M.N. *et al.*, 1999).

Gene and protein sequences can be analysed to identify novel members of existing families or novel families, on the basis of the protein domains the proteins contain. A variety of different protein family databases are available, for example PFAM (see http://www.sanger.ac.u/software/Pfam and Bateman, A., *et al.*, 2002), where the available protein sequence data available in Swissprot, a database of well characterised protein sequences, and TrEMBL, a database of less well characterised translated nucoleotide sequences is analysed for protein domains. In the case of PFAM, sequences from each domain are aligned and a hidden markov model (HMM) is generated which can be used to search for other proteins containing the domain of interest. The proteins can then be characterised on the basis of similarity to these known protein domains. In PFAM, there are currently 3071 protein families which match 69% of proteins in Swissprot and TrEMBL (Bateman, A., *et al.*, 2002).

The ability to compare all known nucleotide and protein sequences with genomic sequence enables novel members of previously characterised gene families to be identified. Genomic sequence can be analysed at the protein level by putative six-frame translation (Gish, W., *et al.*, 1993). This approach will identify those portions of proteins which are similar to previously identified proteins (e.g. specific encoded domains as opposed to whole proteins), but may not identify the entire protein and will not identify the portion of the untranslated region of the gene which encodes the protein. Also, those proteins that are not members of known protein families, or are members of as yet unidentified protein families, will not be identified by comparative protein analysis.

(3) Ab initio gene prediction

The two approaches described above, cDNA sequencing and comparative protein analysis will not identify all the genes in the human genome, and those that are identified are not necessarily going to be complete. Therefore, methods were devised to predict regions of the genome likely to contain genes. These methods took advantage of gene-specific signals such as CpG islands, codon usage and splice sites. CpG islands are associated with approximately 56% of genes (Antequera, F., *et al.*, 1993) and the availability of the genome sequence allows computer programs such as CpGfinder (written by Gos Micklen, The Sanger Institute) to predict the location of CpG islands and using this information infer the possible presence of the 5' end of gene. Using CpG island information alone does not provide any information about the structure of a gene and precludes those genes that are not associated with CpG islands but is a useful strategy that can be used in conjunction with other predictive methods.

Codon usage was first recognised as being useful for gene prediction by Staden, R., *et al.*, in 1982. The triplet codon for each amino is degenerative, but there is a tendency that one particular codon is to code for each amino acid. This leaves a signature of protein coding regions in sequence when compared with non-coding areas. Other signals that are present are acceptor and donor splice sites (often AG and GC respectively – approximately 0.5% of splice sites (Thanaraj, T. A., *et al.*, 2001), translation start sites (commonly ATG), polyadenylation signals (such as AATAAA in 60% of genes) and stop codons (TGA, TAA, TAG).

The first computer prediction programs for gene identification predicted the presence of single exons using primarily codon usage statistics and characteristic sequence signals such as acceptor and donor splice sites e.g GRAIL (Uberbacher, E. C., *et al.*, 1991), and MZEF (Zhang, M. Q., 1997). More recently prediction programs that attempt to predict entire gene structures were developed. These programs (e.g GENSCAN (Burge, C., *et al.*, 1997)) and FGENESH (Solovyev, V. V., *et al.*, 1995)) differ in underlying algorithms used but have the same basic premise: prediction of individual exons based on codon usage and sequence signals, followed by assembly of these putative exons into candidate gene structures. *Ab initio* gene prediction methods are not capable of accurately predicting all the genes in the human genome without overprediction. If the thresholds for prediction were set very low they may capture all genes but with low specificity. In a recent study (Guigo, R., et al., 2000) that looked at the accuracy of gene prediction methods using an artificially generated data set, GENSCAN was shown to accurately predict 90% of coding nucleotides and 70% of the exons were predicted correctly. The study concluded that it is not yet

possible to use computational methods alone to accurately identify the exonic structure of every gene in the human genome (Guigo, R., *et al.*, 2000).

(4) Comparative Genome Analysis

Segments of DNA that have function are more likely to retain their sequence than non-functional segments, as they are under the restraints of natural selection during evolution. Therefore, DNA segments that are conserved between species are more likely to have function. The ideal species to compare with human are those whose form, physiology and behaviour are similar, but the non-functional sequence has diverged sufficiently, thus maximising the possibility to detect the differences between conserved and non-conserved sequences. In practice, there is no single species that provides all the answers for human annotation, as different genes and regulatory regions evolve at different rates. Comparisons with more closely related species will provide information such as gene structures and regulatory elements but these sequences may be obscured by non-functional conservation of sequence and therefore a variety of species, including more distantly related species may be required.

The predicted number of genes in the human genome currently stands at 31,000 (based on draft sequence analysis, see Table 1.3). This is only twice the number needed to make a worm, a fruitfly, or a plant. Sixty percent of predicted human proteins have sequence similarity with proteins predicted from finished genome sequence of other organisms. Sixty-one percent of all fruitfly proteins, 43% of those from worm and 46% of yeast proteins show sequence similarity to predicted human proteins (Rubin, G. M., *et al.*, 2000). Those proteins that appear specific to a

particular species may have either a similar function to proteins from other organisms but the sequence has diverged, or novel species-specific function. More than 90% of the domains in human proteins are present in the worm and the fruitfly. Therefore vertebrate evolution has required the invention of very few domains.

Comparative genome sequence analysis is very informative, and so very accurate tools are required to carry it out. Two main strategies for comparing two sequences are available, local alignment and global alignment. One comparison tool that is capable of comparing sequences is DOTTER and utilises a local alignment strategy (Sonnhammer, E. L., *et al.*, 1995). The concept uses graphical matrix plots where one sequence is drawn on the horizontal axis and the other along the vertical axis of a coordinate system and a dot is drawn where two residues match. Regions of similarity that are co-linear between sequences will result in a diagonal row of dots. Spurious matches give rise to a background of dots. A second example of a commonly-used local alignment tool is BLAST (Basic Local Aligment Search Tool) which measures the local similarity between two sequences (Altschul, S. F., *et al.*, 1990), and has become widely used for searching protein and DNA databases for sequence similarities. An adaptation of BLAST, gapped BLAST was developed that is three times faster than BLAST and is more sensitive for comparing cross-species sequences as it allows for gaps in the alignment and can identify weak similarities (Altschul, S. F., *et al.*, 1997). Where BLAST may have found three similarities between two sequences that individually may not be significant, gapped-BLAST is able to link them together by introducing gaps in the alignment to produce a significant alignment. PIPMAKER is an alignment program that is capable of comparing long sequences (Schwartz, S., *et al.*, 2000). The alignments are generated

using BLASTZ, a derivative of gapped-BLAST, and viewed on a percentage identity plot (pip). An example of a commonly-used global alignment tool to compare two sequences is VISTA (VISualisation Tool for Alignments) which uses GLASS (Global Alignment SyStem) to generate the alignment (Dubchak, I., *et al.*, 2000). Initially a rough alignment map finds the long segments that match exactly and flanking regions that have high similarity. The process is repeated on the intervening regions using successively smaller matching segments. The remaining regions are aligned using standard alignment techniques. Conserved regions are identified by calculating the percentage of identical nucleotides within a window (for example 100 nucleotides) moved in smaller nucleotide increments (for example 25 nucleotides).

Initial strategies using comparative analysis focused on using human and mouse sequence to identify non-coding regions (e.g. (Hardison, R., *et al.*, 1993; Koop, B. F., *et al.*, 1994; Duret, L., *et al.*, 1997; Hardison, R., *et al.*, 1997), but once more substantial data sets were available it became clear that comparing two sequences as closely related as human and mouse was also valuable for protein coding regions (Makalowski, W., *et al.*, 1996; Ansari-Lari, M. A., *et al.*, 1998; Jang, W., *et al.*, 1999). Realistic analysis of the effectiveness of alternative gene prediction methods suggested the need for improved prediction accuracy (Dunham, I., *et al.*, 1999; Guigo, R., *et al.*, 2000). In order to complement human gene identification, the mouse sequencing efforts have been accelerated which has necessitated the need for novel comparison tools (Miller, W., 2001).

Proteins that control gene expression bind to DNA but these binding sites are difficult to predict computationally. Analysing genes that have similar patterns of

expression may aid the process of identification as these genes may be expected to share regulatory elements. Aligning sequences upstream of these genes may identify conservation between similarly expressed genes. DNA containing the stem cell leukaemia (SCL) gene from human, mouse and chicken was sequenced and compared. Regions of conservation between the three species were identified that corresponded with exons of the SCL gene. However other regions of conservation were also shown to coincide with regions of known SCL enhancers. One particular region, (+23 region) was shown using a transgenic xenopus reporter assay demonstrated the region contained a novel neural enhancer (Gottgens, B., *et al.*, 2000).

The identification of control regions using comparative genome sequence analysis has been shown (Brickner, A. G., *et al.*, 1999; Gottgens, B., *et al.*, 2000). Comparative genome sequence analysis alone does not identify control regions, but it does identify conserved regions that are candidates for further experimental investigation.

In order to identify all the features encoded in the human genome a systematic analysis of the sequence is required. In order to identify all genes a combination of large-scale computational analysis, manual interpretation and experimental investigation will be required. Already projects are underway to identify and catalogue as many human genes as possible, such as ENSEMBL (http://www.ensembl.org), UCSC (http://genome.cse.ucsc.edu/cgi-bin/hgGateway) and NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/human). As the DNA

sequence of individual human chromosomes is finished, systematic analysis of the genes and other features provides insights into the features encoded within.

1.4 The human X chromosome

The human X chromosome is estimated to be 164 Mb in size and accounts for one-twentieth of the human genome. Females have two X chromosomes and males have one and the highly degenerate Y chromosome which carries the testis-determining factor. The hemizygous state of the male reveals the phenotypic effect of recessive mutations directly and this, along with X-linked dominant disorders, accounts for large numbers of X-linked genetic diseases, which have a characteristic pattern of inheritance (McKusick, V. A., 1998). Mutation analysis for X-linked disorders in males is easier than for autosomal diseases as they only carry the affected chromosome and this can be analysed directly, whereas a mutation on an autosome is masked by the affected sequence if the individual is heterozygous.

An unique feature of the human X chromosome, shared by mammalian homologues including marsupials, is the inactivation of one of the two chromosomes in females. X chromosome inactivation (XCI) is a mechanism for dosage compensation and was first hypothesied by Mary Lyon in 1961 (Lyon, M. F., 1961). Either the paternally- or the maternally- derived X chromosome is inactivated at random but the same X chromosome is inactivated in future generations of each cell. In the extra-embryonic tissues of mouse and marsupials it is always the paternally-derived X chromosome that is inactivated. The inactive X chromosome replicates late during S phase of

meiosis (Takagi, N., 1974) and is associated with hypoacetylation of the histone proteins H2A, H3 and H4 (Jeppesen, P., *et al.*, 1993; Belyaev, N., *et al.*, 1996). In interphase FISH, the inactive X chromosome appears as a condensed mass or Barr body (Barr, M. L., *et al.*, 1949; Barr, M. L., *et al.*, 1961). In its condensed state, the inactive X chromosome is highly stable and is only reversed in female germ cells at meiosis (Chapman, V. M., 1986).

Initiation of XCI occurs in the early embryo and originates in the X-inactivation centre (XIC) in Xq13, and then propagates along the length of the chromosome (Rastan, S., 1994). The *Xist* (X inactive specific transcript) gene is located within the XIC and is transcribed on the inactive X chromosome only (Borsani, G., *et al.*, 1991; Brockdorff, N., *et al.*, 1991). The *Xist* gene produces a 15-17 kb non-coding mRNA that coats the inactive chromosome (Willard, H. F., 1996; Brockdorff, N., 1998; Panning, B., *et al.*, 1998). The 5' end of *Xist* is essential for initiation of X inactivation, and the 3' end of is essential for chromosome counting – i.e. ensuring one chromosome remains active (Brockdorff, N., 1998, Clerc, P., *et al.*, 1998).

The mechanism for X chromosome inactivation is still largely unknown. Coating of the inactive X chromosome begins at the XIC and spreads to the whole chromosome. Segments of X chromosome without an XIC, through deletion or translocation, remain active. Spreading of inactivation from the X chromosome to regions of translocated autosomal chromosomes occurs but much less efficiently than on the X chromosome portions (Rastan, S., 1983). This observation that all DNA is susceptible to the initial coating of XIST RNA but there is something unique about the X chromosome that promotes the coating led to the hypothesis that interspersed

repeat elements, particularly L1s may play a role (Lyon, M. F., 1998). Recent analysis of the human genome sequence showed that the X chromosome is in fact richer in L1s than any other chromosome, 26% as compared to 13% (Bailey, J. A., *et al.*, 2000).

X inactivation is thought to have arisen early in the evolution of mammals. Monotremes show late replication of part of the X chromosome, which may be a rudimentary form of X inactivation. The transfer of genes from autosomes to the X chromosome and from the X chromosome to autosomes is thought to be limited because of dosage compensation. Therefore genes that are X-linked in one mammal are likely to be X-linked in others (Ohno '67). There are exceptions to this (Graves, J. A., 1996) but it is assumed that these did not disturb the mechanism of dosage compensation. To date, the majority of X-linked genes in humans are shown to be present on the X chromosome in mice. The observation that genes on the short arm of the X chromosome in humans are present on an autosome in monotremes and marsupials suggests that the short arm was of autosomal origin and was added to the X chromosome in eutherian mammals.

There are two blocks on the X chromosome that escape X chromosome inactivation, termed the pseudo-autosomal regions (PARs). These pair with the Y chromosome during meiosis and a varied recombination frequency is observed between males and females. A greater degree of recombination is seen in males due to the obligatory exchange of material within the 2.5 Mb PARs during male meiosis. This higher recombination rate is also seen in other regions of the X chromosome (DMD; Abbs, S., *et al.*, 1990), FRAXA; Richards, R. I., *et al.*, 1991).

There has always been great interest in the X chromosome because of the high proportion of X-linked disorders. It was the first chromosome to have a genetic map, based on RFLPs (Aldridge, J., *et al.*, 1984; Drayna, D., *et al.*, 1985). The most recent genetic map contains 216 polymorphic markers and is estimated to be 216 cM in size (Dib, C., *et al.*, 1996). The published gene map (Deloukas, P., *et al.*, 1998) (updated electronically in 1999 – http://www.ncbi.nlm.nih.gov/GeneMap99) indicated the X chromosome may be relatively gene poor, containing only 50% of the number of ESTs as those expected based on its size and assuming a random distribution. It is not known whether the relative paucity of genes on the X chromosome is due to its role in sex determination or the mechanism arose independently. The intense interest in the X chromosome led to rapid progression of the YAC map and successive X chromosome workshops produced consensus landmark maps on regional efforts. This culminated at the most recent workshop in the generation of YAC maps covering virtually the entire X chromosome (7^{th} X chromosome Workshop (7XCW) – The Sanger Centre 1995; Nagaraja, R., *et al.*, 1997).

Also at the 7XCW, responsibility for bacterial clone mapping and sequencing of the X chromosome was divided up between centres. This underlies the international collaborative effort involved in sequencing the human X chromosome. Currently there are 26 bacterial clone contigs covering the X chromosome and 125 Mb of finished sequence and 65 Mb of draft sequence is available (see http://www.sanger.ac.uk/ChrX). Efforts are continuing to close the remaining gaps and finish the sequence.

1.4.1 Xq22

The region in Xq22 under study in chapter 3, between DXS366 and DXS1230 encompasses approximately 7 Mb on the long arm of the X chromosome. This region has been shown to contain a number of genes involved in genetic disorders, some of which have already been isolated. These include the genes involved in Fabry disease (Bernstein, H. S., *et al.*, 1989), Pelizaeus Merzbacher disease (Hudson, L. D., *et al.*, 1989; Trofatter, J. A., *et al.*, 1989), and X-linked agammaglobulinaemia (Tsukada, S., *et al.*, 1993; Vetrie, D., 1993). Other disease loci mapping to the region, for which candidate genes have not yet been identified include X-linked megalocornea (Chen, J. D., *et al.*, 1989) and X-linked deafness 2 (DFN2; Tyson, J., *et al.*, 1996). The majority of the region between DXS366 and DXS1230 is thought to lie within a light band hence is expected to contain many genes (discussed earlier, see Section 1.2). At the start of this work, physical mapping had concentrated at the resolution of the YAC map (Vetrie, D., *et al.*, 1994; Kendall, E., *et al.*, 1997). The generation of sequence-ready bacterial clone contigs covering the region, described in chapter 3, will provide the first step towards the complete genomic sequence, and a fully

annotated version containing all the genes and other biologically relevant information.

1.4.2. Xq23-24

The region in Xq23-24 under study in chapter 4, between DXS7598 and DXS7333 encompasses approximately 8 Mb of the long arm of the X chromosome. This region has been shown to contain a number of genes including ANT2 (Nagaraja, R., *et al.*, 1998, Schiebel, K., *et al.*, 1994, Steingruber, H. E., *et al.*, 1999) and LAMP2 (Manoni, M., *et al.*, 1991, Nagaraja, R., *et al.*, 1998, Steingruber, H. E., *et al.*, 1999). Xq23 is a dark staining R-band and is expected to contain few genes whereas Xq24 is a light staining G-band and is thought to be gene rich. Initial evidence that Xq24 is gene rich was shown by a cluster of CpG islands mapping to the region (Maestrini, E., *et al.*, 1990). The production of sequence-ready maps and the generation of genomic sequence will enable a systematic approach for gene identification to be undertaken.

1.4.3 Non-specific X-linked mental retardation

One of the diseases whose critical region is contained within Xq23-24, MRX23 (Gregg, R. G., *et al.*, 1996), is one of many non-specific mental retardation (NSMR) disorders mapping to the human X chromosome. NSMR includes all those disorders whose only consistent clinical manifestation is mental retardation and includes X-

linked mental retardation (XLMR) (Neri, G., *et al.*, 1999). It has been known for a long time that there is an excess (25-30%) of males among the mental retardation patients, particularly with a mild to moderate phenotype (Lehrke, R. G., 1974).

XLMR represents approximately 5% of all mental retardation and corresponds to a prevalence of 1 in 600 males in the general population (Crow, Y. J., *et al.*, 1998). Regional assignment along the X chromosome of different families with XLMR has shown that at least fifty MRX families exist (Toniolo, D., *et al.*, 2000; Neri, G., *et al.*, 1999). By convention, each family represents a locus and is designated by the acronym MRX and by a progressive number. A database listing XLMR disorders has been developed where the XLMR have been divided into two groups, the syndromic and the non-specific XLMR (Cabezas, D. A., *et al.*, 1999).

Eight genes for NSMR have been identified to date and form a heterogenous group encoding diverse proteins ranging from transmembrane proteins to transcription factors (Toniolo, D., *et al.*, 2000). However, all the genes identified so far are either directly or indirectly involved in signalling pathways. Further study of these genes and the identification of more NSMR genes are required before a full understanding of NSMR and the development of cognitive function are achieved.

1.5 Aims of this thesis

At the time this thesis began, the majority of physical maps covering large portions of the human genome were in the form of YAC maps. However, the plans to generate human genome sequence necessitated the construction of physical maps in

bacterial clones, a more suitable substrate for sequencing. The aim of the first part of this thesis was to generate sequence-ready bacterial clone contigs across a 7 Mb portion of the long arm of the X chromosome in Xq22.

The generation of the sequence of the human genome is only the first step in its complete characterisation. An important subsequent step is the identification of all the genes and other biologically relevant information encoded within. The second aim of this thesis was to construct a transcript map in Xq23-24, identifying and experimentally confirming as many genes as possible using the resources available at the time. Gene identification using sequence similarity searches in combination with *ab initio* gene prediction to predict genes that are then confirmed by cDNA isolation and sequence has two major limitations: not all genes will be identified by this method and not all of those predicted genes will be confirmed experimentally.

Comparative genome analysis is playing an important role in the elucidation of all the features within the human genome and in the understanding of their function. Generating DNA sequence from syntenic portions in other species allows functionally conserved units to be identified at the sequence level. This analysis provides additional data that can be used to support previously identified genes as well as identifying potential novel functional units. The final aim of this thesis was to construct bacterial clone contigs in mouse and zebrafish, sytnenic to a region in human Xq24. The sequence from all three species will allow further analysis of the features previously identified in the human sequence and identify potential novel functional units.

Chapter 2

Materials and Methods

Materials

2.1 Chemical reagents

All common chemicals were purchased from Sigma Chemical Co., BDH Chemical Ltd., and Difco Laboratories unless specified below or in the text.

Amersham Pharmacia Biotech	Dextran sulphate, Na^+ salt
Bio-Rad Laboratories	β-mercaptoethanol
Gibco BRL Life Technologies	Foetal bovine serum
	ultraPURE™ Ammonium sulphate, enzyme grade
	ultraPURE™ agarose
Roche Applied Science	Restriction Buffer B
Stratagene®	Perfect Match® (1 U/µl)
	Taq Extender

2.2 Enzymes and commercially prepared kits

All restriction endonucleases were purchased from New England Biolabs, unless listed.

Amersham Pharmacia Biotech	T4 DNA ligase (1 U/µl)
	*Sau*3A1
Bio101Inc	Geneclean II
Gibco BRL Life Technologies	M-MLV reverse transcriptase
New England Biolabs	T4 DNA ligase
PE Applied Biosystems	Amplitaq™
	TaqFS
Qiagen	DNA gel purification
Roche Applied Science	Klenow enzyme (sequencing grade, 5 U/µl)
	T4 Polynucleotide kinase
Sigma	Ribonuclease A

2.3 Nucleotides

Amersham Pharmacia Biotech	Redivue™[α-^{32}P]-dCTP (AA 005) aqueous solution (370 Mbq/ml, 10 mCi/ml)

	Redivue™[γ-^{32}P]-dATP (AG 1001) aqueous solution (370 Mbq/ml, 10 mCi/ml)
	[α-^{35}S]-dATP (Q11135) (370 Mbq/ml, 400 Ci/mmol)
PE Applied Biosystems	Fluorescently labelled (TET, HEX, NED) dideoxyadenosine triphosphate (ddA)
	Fluorescently labelled (ROX) dideoxythymidine triphosphate (ddT)
Amersham Pharmacia Biotech	2'-deoxynucleoside 5'-triphosphates (dATP, dTTP, dGTP, dCTP)
	dideoxyguanine 5' -triphosphate (ddGTP)
	Random hexanucleotides pd$(N)_6$, 5'-PO_4, Na^+ salt

2.4 Solutions

Solutions used in the present study are listed below, alphabetically within each section. Final concentrations of reagents are given for most solutions. Amounts and/or volumes used in preparing solutions are given in some cases. Unless otherwise specified, solutions were made up in nanopure water.

2.4.1 Buffers

10x Ligase buffer	500 mM Tris-HCl (pH 7.4)
	100 mM dithiothreitol
	100 mM $MgCl_2$
10x PCR buffer	670 mM Tris-HCl (pH7.4)
	166 mM $(NH4)_2SO_4$
	67 mM $MgCl_2$
1x TE	10 mM Tris-HCl (pH 7.4)
	1 mM EDTA
1x $T_{0.1}E$	10 mM Tris-HCl (pH 8.0)
	0.1 mM EDTA

2.4.2 Electrophoresis and hybridisation solutions

6x Buffer II	0.25% bromophenol blue 0.25% xylene cyanol 15% ficoll
Denaturation solution	0.5 M NaOH 1.5 M NaCl
Formamide dyes	80% v/v deionised formamide 0.1% w/v bromophenol blue 0.1% w/v xylene cyanol 1 mM EDTA 50 mM Tris-borate (pH 8.3) (*i.e* 0.56x TBE)
Formamide dyes mix	0.0075% w/v SDS 3.75 mM EDTA 1.6x formamide dyes
6x Glycerol dyes	30% v/v glycerol 0.1% w/v bromophenol blue 0.1% w/v xylene cyanol 5 mM EDTA (pH 7.5)
Neutralisation solution	1.5 M NaCl 1 M Tris-HCl (pH 7.4)
20x SSC	3 M NaCl 0.3 M Trisodium citrate
10x TAE	400 mM Tris-acetate 20 mM EDTA (pH8.0)
10x TBE	890 mM Tris base 890 mM Borate 20 mM EDTA (pH 8.0)

2.4.3 Media

All media were made up in nanopure water and either autoclaved or filter-sterilised prior to use.

For agar used for bacterial growth 15 mg/ml bacto-agar was added to the appropriate media.

Antibiotics were added to media as appropriate (see Table 2.1) to the following final concentrations: Ampicillin (sodium salt dissolved in 1 M sodium bicarbonate, stored at -20°C), 100 μg/ml; Tetracycline (dissolved in absolute ethanol, stored foil-wrapped at 4 °C), 5 μg/ml; Kanamycin (purchased as a solution, stored at 4 °C), 30 μg/ml; Chloramphenicol (stored at 4 °C), 12.5 μg/ml (all supplied by Sigma).

Table 2.1: *Clones and appropriate antibiotics*

Clone type	**Library**	**Antibiotic**
Plasmid	NA	Ampicillin or Tetracycline
Cosmid	LL0XNC01	Kanamycin
PAC	RPCI1,3,4,5, 6	Kanamycin
BAC	RPCI-11, 13	Chloramphenicol

LB

10 mg/ml bacto-tryptone
5 mg/ml yeast extract
10 mg/ml NaCl
(pH 7.4)

2 X TY

15 mg/ml bacto-tryptone
10 mg/ml yeast extract
5 mg/ml NaCl
(pH 7.4)

2.4.4 DNA labelling and hybridisation solutions

100x Denhardt's	20 mg/ml Ficoll 400-DL
	20 mg/ml polyvinylpyrrolidine 40
	20 mg/ml BSA (pentax fraction V)
Hybridisation buffer	6x SSC
	1% w/v N-lauroyl-sarcosine
	10x Denhardt's
	50 mM Tris-HCl (pH 7.4)
	10% w/v dextran sulphate
OLB3	240 mM Tris-HCl (pH 8.0)
	75 mM β-mercaptoethanol
	0.1 mM dATP
	0.1 mM dGTP
	0.1 mM dTTP
	1 M HEPES (pH 6.6)
	0.1 mg/ml hexadeoxyribonucleotides (2.1 OD units/ml)

2.4.5 General DNA preparation solutions

GTE	50 mM glucose
	1 mM EDTA
	25 mM Tris-HCl (pH 8.0)
3 M K^+/5 M Ac^-	60 ml 5 M potassium acetate (pH 4.8)
	11.5 ml glacial acetic acid
	28.5 ml H_2O

2.5 Size markers

1 kb ladder (1 mg/ml) (Gibco BRL Life Technologies)
Contains 1 to 12 repeats of a 1,018 bp fragment and vector fragments from 75 to 1,636 bp to produce the following sized fragments in bp: 75, 142, 154, 200, 220, 298, 344, 394, 516/506, 1,018, 1,635, 2,036, 3,054, 4,072, 5,090, 6,108, 7,125, 8,144, 9,162, 10,180, 11,198, 12,216.

Lambda DNA/*Hin*d III (Gibco BRL Life Technologies)
Contains *Hin*d III restricted dsDNA fragments of the following sizes (kb): 23.13, 9.416, 6.557, 4.361, 2.322, 2.027, 0.564, 0.125

Analytical marker DNA wide range (Promega)
Provides an evenly spaced distribution of DNA fragments from 0.702 kb to 29.95 kb

DNA molecular weight marker V (Roche Applied Science)

2.6 Hybridisation membranes and X-ray and photographic film

Amersham	Hybond-N™ Nylon (78 mm x 119 mm) (used for high-density clone gridding)
Polaroid	Polaroid 667 Professional film
Autoradiographs	Fuji RX medical X-ray film

2.7 Sources of genomic DNA

Human placental DNA for pre-reassociation (ready-sheared) was purchased from Sigma Chemical Co.. Human placental DNA for PCR was purchased from Sigma Chemical Co. DNA from hybrid Clone 2D (Cl2D) that contains the entire X chromosome was kindly provided by Adam Whittacker. DNA from two affected individuals from family MRX23 was kindly provided by Ron Gregg. DNA from a normal male control sample was kindly provided by Alison Coffey.

2.8 Bacterial clone libraries

2.8.1 Cosmid libraries

Cosmids from the Lawrence Livermore flow-sorted X chromosome cosmid library (LL0XNC01) (prefixed 'cU') were kindly provided by Dave Vetrie and Elaine Kendall. Cosmids from a library constructed from a male with 5 X chromosomes (Holland, J., *et al.*, 1993) (prefixed 'cV') were also kindly provided by Dave Vetrie and Elaine Kendall.

2.8.2 PAC and BAC libraries

The RPCI-1, RPCI-3, RPCI-4, RPCI-5 (prefixed 'dJ'), and RPCI-6 (prefixed 'dA') PAC libraries, and the RPCI-11 (prefixed 'bA') and RPCI-13 (prefixed 'bB') BAC libraries were used as a source of human derived PAC clones and BAC clones respectively in this thesis. Mouse-derived BAC clones were obtained from the RPCI-23 (prefixed 'bM') library, and zebrafish-derived BAC clones were obtained from the RPCI-71 libraries (prefixed 'bZ'). These libraries were all kindly provided by Pieter de Jong and Joe Catanese (see http://bacpac.med.buffalo.edu/), and imported and maintained by the Sanger Institute Clone Resources Group.

2.8.3 cDNA libraries

A range of up to 20 different cDNA libraries were used in this study (see Table 2.2). cDNA libraries were imported and maintained by Jacqueline Bye and Susan Rhodes. Each library contains 500,000 cDNA clones, divided into 25 pools of 25,000 clones. Five pools were combined to form a superpool containing 100,000 clones. Prior to their use in PCR, each superpool was diluted 1:100 and 1:1000 in $T_{0.1}E$.

Table 2.2: *cDNA libraries used*

cDNA library code	cDNA library description	Vector	Source/ Reference
1. U	(Monocyte NOT activated-from a patient with promonocytic leukaemia) (U937+)	pCDM8	Simmons (1993)
2. H*	Placental, full term normal pregnancy (H9)	pH3M	Simmons (1993)
3. P	Adult brain	pCDNA1	Pfizer
4. DAU	B lymphoma (Daudi)	pH3M	Simmons (1993)
5. FB	Fetal brain	pCDNA1	Invitrogen
6. FL	Fetal liver	pcDNA1	Invitrogen
7. HL	Peripheral blood (HL60)	pCDNA1	Invitrogen
8. SK	neuroblastoma cells	pCDNA1	Invitrogen
9. T	Testis	pCDM8	Clontech
10. FLU	Fetal lung	pCDNA1	Invitrogen
11. AL	Adult lung	pCDNA1	Clontech
12. UACT*	(Monocyte PMA activated - from a patient with promonocytic leukaemia) (U937act)	pCDM8	Simmons (1993)
13. YT*	HTLV-1+ve adult leukaemia T cell	pH3M	Simmons (1993)
14. NK*	Natural killer cell	pH3M	Simmons (1993)
15. HPB*	T cell from a patient with acute lymphocytic leukaemia (HPBALL)	pH3M	Simmons (1993)
16. BM*	Bone Marrow	pH3M	Simmons (1993)
17. DX3*	Melanoma	pH3M	Simmons (1993)
18. AH	Adult Heart	pcDNA3-Uni	Invitrogen
19. SI **	Small Intestine	pcDNA3	Stammers

* Generously provided by Dr Simmons, Oxford (Simmons, D., *et al.*, 1993).

** Generously provided by Dr Stammers (Sanger Institute)

2.9 Primer sequences

All primers were synthesised in house by Dave Fraser or externally by Genset. Table 2.3 lists the vector-specific primers and sequences used in the vectorette method. Table 2.4, 2.5, 2.6, 2.7, 2.8 and 2.9 list the STSs used in this thesis, the sequence and size in base pairs (bp) of each primer, and the optimal annealing temperature (AT – given in °C). Where appropriate, the clones, or genes from which the STSs were derived are also listed.

Table 2.3: *Vector-specific primer sequences and 'bubble' sequences for primers used in vectorette PCR and SSPCR (performed on clone DNA and cDNA)*

Primer Name	Primer Sequence
SP6PAC*	ATTTAGGTGACACTATAG
T7PAC*	TAATACGACTCACTATAGGGAGA
PACS2*	GCTAGGAGGGCTTAACTGAT
PACT2*	CTGGGTTGAAGGCTCTCAAG
224**	CGAATCGTAACCGTTCGTACGAGAATCGCT
BPHI**	CAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGGACGACAG AAGGGAGAG
BPHII**	CTCTCCCTTCTCGAATCGTAACCGTTCGTACGAGAATCGCTGTCCT CTCCTTG
pH3M1FP	CTTCTAGAGATCCCTCGA
pH3M2FP	GCTCGGATCCACTAGTAA
pH3M1RP	CTCTAGATGCATGCTCGA
pH3M2RP	CGACCTGCAGGCGCAGAA
pCDM8.RP	TAAGGTTCCTTCAGAAAG
T7.2FP	AATACGACTCACTATAG

* designed by John Collins (Sanger Institute)

** (Riley, J., *et al.*, 1990)

Table 2.4: *STSs from Srivastava, A.K., et al., (1999) used for contig construction as described in chapter 3*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)
sWXD797	GCCTTGGAATATCTTCCTAC	AAACATTTGTGAGTCATCAGTGTC	83	55
sWXD940	CATGCATAATGCATAGCATGG	TGGTAAGAGCTTAAATTTGCTAAGGG	65	60
sWXD1199	GCGAAGAAAACATTACCTGG	CAGGATATCAAAAAACCTCAACTG	71	60
sWXD1259	GGGAAGAAATGAAAGGAGG	AGTCAGTCCCACTCTTGTC	65	60
sWXD1283	TTGGAGTAGACAACAGGAC	TAAATTAAAGGGAAGCACTAAGAG	140	60
sWXD1440	CCCTGCTCCTACTCATAGC	GGGCTTGGTAGTAGTGCTTG	100	60
sWXD1549	CAAACAGAATGATTAAATACAGGAC	GTAGGAGGCTACTAAGAAG	126	60
sWXD1805	CATATTTTGTCAGTGTGGTC	GCCTTAACTATAAGAACCAG	140	60
sWXD2782	CCCATTCCAGATATAGATTATCAG	TTTCATCACAACATCCCCAAAGAC	84	60
sWXD2783	GTGATGAGAGCTTAAAAATCCC	ACCAGCATTCAAGAAAGGTGAG	84	60
sWXD2789	ATTCCTTACCCACACAGTC	AATATAACGTGCATGTATGGTTTC	140	60
sWXD2804	CAATTCACATAGTCCTCAAACC	TCTCCTACAGATGATTAACCACC	160	60
sWXD2841	GGGAACAAGGAAGAAAGAATG	TCCCTAAGACCCTCCTATG	81	60
sWXD2843	GTGGGTCCTACTTTCAGGG	CTGACGAGTTCATGCAAGCC	69	60
sWXD2844	GAAGGTTTAGGTGGGATAATG	CATCTCTTGTCTGGACTATTTC	145	60
sWXD3375	TAGTTTAATGGAATGGTAGAATGG	TTCACATACAGCTTCCTGG	97	60
sWXD3381	GAAGTAAGATGGGACAATAGG	CAGTTTTACGGAAATGAACAGTAG	107	60

Table 2.5: *STSs derived from clone ends used for walking as described in chapter 3*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)	Clone Name
stSG14873	GAAGATTGTTGCCAGAACTGC	GCCCAGCCAAAGAATTACAA	103	60	cU101D3
stSG14874	AATGGCCAGTAAAGACAGAAGA	ATGACTCAATGCCTAAAAAGGA	82	60	cU235H3
stSG14875	AATTGGCGATCATAGCTTTAGC	CTCTGGAATCACAAATGTGGG	104	60	dJ127B14
stSG14875	AATTGGCGATCATAGCTTTAGC	CTCTGGAATCACAAATGTGGG	104	60	dJ127B15
stSG14929	AAGGAAAAGCAAGAGAAGGACA	AGTGTGCATTGCATAGCTGG	118	60	cU105G4
stSG14930	AGCCGTGTTAATCTTGACACTC	TGGGGGAAAAACATTCATGT	104	60	cU116E9
stSG14933	CATGCTCATTTTAACACTTGCC	AACCCCTTTCCTAAGTAGTGGC	101	60	cU160A4
stSG14958	ATTTTACATGTCCAGGACAGGG	TCAAAAAGAACACCGCACC	119	60	cU105G4
stSG14962	ACACTGAAGCCTTTTGGAGG	TCATGGGGGTTTGTTGTACA	118	60	cU17A7
stSG14963	GTGTGTGTGTATTTAACAGGCG	GGCAGTTGTCAGCTAAATAGCC	183	60	cU212C1
stSG17548	GCACAGTGCTTGGCACAC	CTCCCTCAGGTACACTGGTAAG	127	60	cU235H3
stSG17555	ATTTATTGAGTTGGCATCCCC	TTTCCCCCCATACTGTCAAA	121	60	dJ334P19
stSG17563	AAAGGATGAAATGACTCTTGCC	TACCACCAGTTTAGCAGGCC	154	60	dJ82J11
stSG17563	AAAGGATGAAATGACTCTTGCC	TACCACCAGTTTAGCAGGCC	154	60	dJ334P19
stSG22771	GGTTGTAGGTGTGCATGTGC	GCAAAGCGTTCTGAATACCC	121	60	cU159B9
stSG22771	GGTTGTAGGTGTGCATGTGC	GCAAAGCGTTCTGAATACCC	121	65	cU159B9
stSG22772	TGCTAGCACAACAGGGTGAC	ATCATGGAGAATGGGGTATCC	135	60	cU159B9
stSG22773	TTGACAGCATAATCCACTTTGG	TGGTTCTTTCAGCATCTGTCA	181	60	cU160A4
stSG22774	CTCGCTTTTCCTTTTTGGC	TTTTATTACCCAATCAGCCCC	149	60	cU50F11
stSG22775	AAGGCCTTACCATTGTCCCT	TTTTTCTTGGGCAATTCCAG	170	60	cV602D8
stSG27821	AGCAATCCCACAGCTAGGC	TGTTGATGGACACTTAGGTTGC	133	60	cU232G2
stSG27821	AGCAATCCCACAGCTAGGC	TGTTGATGGACACTTAGGTTGC	133	60	cU232G2
stSG27822	TTTTTTTTTTGACGGAATCTCA	TGGTGGTGTGCACCTGTAGT	145	60	cU232G2
stSG38412	GCAAAAATAAATGGTTGGAAGG	TCCCAGAGGTAACCGTTATCC	101	60	dJ79P11
stSG41236	TTGAGACCTGAATAGCTCCCA	TATTGCTGAAACCACTTTGGG	167	65	dJ148H18
stSG41239	GCGAGTGGTGCAAGAGTGT	GTCAGGAGAGTGTGTGAAATGG	165	65	dJ306P24
stSG41240	TCCTAAGGGCCTTGATGATG	GAGATACTGGACAGCTATGGGG	138	65	dJ34L20
stSG45649	TGTGTTGCTTTTGCTTCCTG	AAGTCCTTACAGTCAGCAAGGC	140	60	cU35G3

stSG45650	TGGAGGGTCATAAGGCAAAG	TTGAACCTCTCAAGTCGCCT	140	60	cU35G3
stSG45651	CTGCAAAACTGGCAGTACCA	CACTGACCAGTGATTTTCAAGG	123	60	cV362H12
stSG45652	AGAACATGGGGTTCTTGGG	TTCACAAAACCAATAAAAGCCC	130	60	cV461C10
stSG45653	TTGCCTGATCATATGAATCACC	CACTCACGTTGCTGTGGG	141	60	cV461C10
stSG45654	AGCCTTTTCAGATTCTGAGCC	TGTCAGCAGAGTGTCATCCTG	137	60	cV698D2
stSG45655	TAGTGACATGTGGAAATGCCA	GGCATGGCTCTTTGTCCTAA	182	60	cV698D2
stSG45656	TCATTTTGACTTTGTGGAGGC	AGGAAAGCCCAGAAGAAAGG	163	60	dJ409J21
stSG45751	GTATGCCATTTCCAATCAGATG	TTGCCCCAGATTTGCTTC	86	60	cU157D4
stSG45752	CTTGGCACAAAGAATGGGAT	AGAAGATGGGTTTCTGGGCT	129	60	cU157D4
stSG45753	AGCTCCAAGCCAACTTGAAA	GTTCACTTAAAGGGTGGAGCC	120	60	cU19D8
stSG45754	CTGAATCTTTGCATGATTGCA	CATTCTAAACATGTGCTCAGGG	177	60	cU19D8
stSG45755	ACTATGGATTCTGTCCCCAGG	TCCAGGCTACTACCCAAATCC	197	60	cU237H1
stSG45756	AGCGCCATTAGATGAGCAAT	AGATGGCCCCTTCCTCTTAA	213	60	cU237H1
stSG45757	CTGGCTCACATTCAGGGC	TGTTAAAAACAACCCGCTCC	174	60	cU46H11
stSG45758	GTGCATCCTATGAAGCACATG	ATCTGCAGGTGATTCCTGTACA	133	60	cV351F8
stSG45759	TGAGCCACCGCTAATAAAGG	GAATAGGACAGCCCTTTCCC	180	60	cV351F8
stSG45761	TGAGAACCACTGCATCATAAGG	GTTTTCCCTTTTGAACTTGCC	102	60	cV389H8
stSG45762	TTCTGTCATTTGGGACACCA	TGGTTGTGTGTTTTTTAGGCA	135	60	cV521F8
stSG45763	GCCATTTTCACTTATTGTGGTT	GAGCAATAAAGGAAAAAATGCA	146	60	cV602D8
stSG45764	GGCAAAGTTCAGCTCAGGAC	CTAGGAAGTGCTTTGGCTGG	180	60	cU65A4
stSG49025	CAGCAGTCTTCTAGGTGCCC	TAATACCCAGCTGTTGGAACG	155	60	cU86H4
stSG49026	TGGGAAATGCTCCTCTGGTA	TTTCCGTGTCTTGGGAAAAG	122	60	cV618H1
stSG49027	GCCCAGAAGGTGTTAACTTCC	CAATGATGGCATTTCATATTGC	121	60	dJ1184O6
stSG49032	AGGGGAGAGAACAGCACTAGC	TGGGAAGGCACTAACATTCC	147	60	dJ198P4
stSG49032	AGGGGAGAGAACAGCACTAGC	TGGGAAGGCACTAACATTCC	147	60	dJ198P4
stSG49038	GTGCTTCCATAGCTTCATCTCC	CACAAAGGTTAGAGCACACAGC	123	60	dJ738A13
stSG49038	GTGCTTCCATAGCTTCATCTCC	CACAAAGGTTAGAGCACACAGC	123	60	dJ738A13
stSG50178	ATGGCAAACAGAGAGCTGGT	ACTGTGGCTGCAGGTTCTTT	170	60	cU65A4
stSG61723	ATCTGATCATTCTGGCCCAG	GATTCAGCCAAGCACATGAA	122	60	cU165H7
stSG61725	TCAACAGGGAACAACTTGACC	GGACTGTCTCTTCAAAGTTGCC	122	60	cU96H1
stSG61726	TAGAAGGGCCTCATGTGTCC	TGGGCAAAATGTCCCAAC	145	60	cV870H8

stbA191C22SP6	CACCATCACAATGCATACTGC	GGCAATTTGTTAGTATTTGGCA	140	60	bA191C22
stcU105G4.1	GCAGCTGTTTTTGCTAAGGG	AAAGCTGGTTTGTTCTCTCTGC	139	60	cU105G4
stcU105G4.1	GCAGCTGTTTTTGCTAAGGG	AAAGCTGGTTTGTTCTCTCTGC	139	60	cU105G4
stcU173H7SP6	GAGCTCCTTCTGATCTTGGTC	AGTTCCACATGGTCAAAGCC	99	65	cU173H7
stcU212C1T7B	AAGCTACTTTGAGTGCTTTGGC	AGTGTGGACAACATCTGGAGG	133	60	cU212C1
stdJ77O19.1	CAATGGGGTGCTAGTGGAGT	CTAGCATTTCCCAAGACCCA	167	60	dJ77O19
stdJ82J11	CTCCTTCAGATGCAATTGATTG	GAGGGTGTTCATTCAAAAAAGG	186	60	dJ82J11
stdJ148H18T7A	AGGGATCAGCAACATTGACC	AAAACAATTGCATCGAAGGG	172	60	dJ148H18
stdJ233G16T7	CTTTCCATTTCTACCGTCATCC	GTTCCGATTTAGGCTTTCAGG	158	60	dJ233G16
stdJ258H17SP6.224	CGTTTCAAAGTCCATGGGTT	GCCATTTAGAACCTCTGCCA	174	60	dJ258H17
stdJ258H17T7.224	AAAAAAAATTTTCTGCTGGTGG	AAATAGGCCTGCTCGTTCAA	122	60	dJ258H17
stdJ324P21SP6	AACTCCAGGTTCTGTAGCAAGC	GTACTGGCCCGGTTTACTAGC	177	60	dJ324P21
stdJ324P21SP6	AACTCCAGGTTCTGTAGCAAGC	GTACTGGCCCGGTTTACTAGC	177	60	dJ324P21i
stdJ341D10T7	TGCATTGGGTGCAAAATTTA	GCCTCAGTGAGCCTTACCTG	168	60	dJ341D10
stdJ400D4T7	CTGTTCCTCAAAGCTTGTAGCA	TAAGTACTGTGATGGGCATTGG	150	60	dJ400D4
stdJ421I20SP6	GGAAAGGAGAAGAAAAGGCC	TCTGTGCCTGCAACCATG	122	60	dJ421I20
stdJ479J7T7A	GACCACCTGGCCTAACTTCC	CCAAATTAGGAAAGACTCCATG	121	60	dJ479J7
stdJ479J7T7A	GACCACCTGGCCTAACTTCC	CCAAATTAGGAAAGACTCCATG	121	65	dJ479J7
stdJ519L22T7.224	TTCAAAGATTGGCAAGATTGG	GTGGATCCTTGAAAAACAAAGC	130	60	dJ519L22
stdJ663P11SP6	CGCAGTTTACTTAAGGGGACC	GGAACCTAAGGAGTGGGCTT	137	60	dJ663P11
stdJ664K17T7	GTATCAGAGGCCAAGCTCATG	GGAAGAAGTGCTGATGAGGG	142	60	dJ664K17
stdJ777L12T7	GTTTCCTGGCAGAAGCAGTC	CTTTCATCAGGGTGAACTGTTT	147	60	dJ777L12
stdJ823F3SP6	AACTGCTTTGTTAATGCCTGC	GCTTTACATGTGAGTGCTCAGG	123	60	dJ823F3
stdJ969N12T7	GCCTTTCATAATTTCTTCCAGC	TGGAATAAATGCTTGAAATTGC	126	60	dJ969N12

Table 2.6: *STSs used for gene identification as described in chapter 4.STSs are named with the prefix 'st' followed by the clone name from which the STSs was derived. The number distinguishes multiple STSs designed in the same clone. The 'n' indicates nested primers for SSPCR. These were not used as primer pairs and so no annealing temperature is given.*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)
stbA45J1.1.1	CTCTCAGCTCTCGGAAGGAC	TGCTGAGTCAGGGACTGATG	142	60
stbB125M24.1.1	TTGCAAGCATCACTTCTTGG	GCTGGTATCTTGTGTCAAATGC	140	60
stbK421I3.1	GCCCAGGACTCTTCTTCCTC	GGGATACTGAGAGCATCGGA	123	55
stbK421I3.2	GAGAACCAGAAGGGCGGT	AATGCTGTCTAGCTCCTTCAGG	162	60
stbK421I3.2n	TGGCCACCAGGAGCCCTG	GCTGGGGCTGAATAGACG	-	-
stbK421I3.3	TTCCAGCAGCCTGTGTTTC	ATAACAAAAGGGAATGGGCC	123	65
stbK421I3.3n	CACCTATGCCACCCGCTG	GCATGAGTGGAAGGGGCAAG	-	-
stdA39H21.1	TGTGCTGGTTCTGGCAGC	TTGTTGACTGAGGCAGATAAGC	124	60
stdA155F9.1	GTGACGAATCCACATCCTG	GTTCTGCACAGTGTGTAATG	82	60
stdA155F9.1	GTGACGAATCCACATCCTG	GTTCTGCACAGTGTGTAATG	82	60
stdA155F9.1n	GCCAGAATTGAAAAGGTAC	GTACCTTTTCAATTCTGGC	-	-
stdA155F9.2	CAAAGTTGTTGAGCCCCTG	GATATATCTTCCATTGGGAAC	108	60
stdA155F9.2n	GACTATGAGAATGTTATTG	CTTTGGGCAATAACATTCTC	-	-
stdJ29I24.1	CTCCGGCTCAGTCTTACAGG	AAGTTTGCTAGCCACGCG	127	60
stdJ57A13.1	TGTGATAGAAGCACGCAAGG	TATTCACCAAAATCAGCTGTGG	160	60
stdJ57A13.2	GAGCCCACTTTGGTGGTG	GTAAAGGGAGAAGTGCAACCC	129	60
stdJ57A13.2n	GCAAGGCCTGGCTGGGTTCT	TTATTACCTCCAGCACAGGA	-	-
stdJ93I3.1	TCCTGAAGACAGCTGCCC	TGGTTTTTCCTCCAATTTTTC	87	60
stdJ93I3.2	TCTTTGCAGCTGTGGCTCTA	CACCCAGTTGATGTGACAGG	152	60
stdJ169K13.1	GACCACTTCACCCTGTCCTG	GCCGCAGTAGCTCAGCTC	120	60
stdJ169K13.2	AGGACATGGAGTTCACCGAG	CTAGGCCATCTCCTCCTCG	124	60
stdJ169K13.3	AGAAGGTACTGCCTCAGTCTGG	CCCTGGATCTGTCTCCAGAA	127	60
stdJ170D19.1	GATGATGGATACCCAGTGGC	TCATCATCTACCACTGGGCA	92	60
stdJ170D19.1n	GATGATGGATGCCCAGTG	ATCCATCATCTGCCACTG	-	-
stdJ170D19.2	CAAAAATGGAGCTTTGTCAGC	TCAAAAGAAAAGCGCATGC	163	60

stdJ222H5.1	CTCCAAGATTCAACTATGTGGG	TCCCAAACAACTCAAGCTCC	125	60
stdJ222H5.2	TCCTTGCCCATGCAAATC	GATCCCATGGAACTGAAGGA	119	60
stdJ278D1.1	TCACTAGCAGATGCCATCATG	ACCAGCTTCGACTTGAAGGA	103	60
stdJ278D1.1n	TCTATGCACAGGGAAAGC	GTGATGTGACCCTGATGC	-	-
stdJ318C15.1	ATTTTAGGGACATGGGACTGG	CAGACAGCATGCTTAAAAGGC	161	60
stdJ318C15.1n	GAGTTATTTAGGGCTCATATT	TTCGAAGTAAACTTCTATCA	-	-
stdJ321E8.1	CTCATACCTGCCTCCTGCTC	TCCAGTCAGATGGAGATTTGG	128	60
stdJ321E8.2.1	CGAAAAGTGGGATGAAGAGG	TGGATTTTCTTGGCTTCACC	136	60
stdJ321E8.3.1	ATGCCTGTGGGAATTGTAAC	TTGAACACTGTCACATACATCCA	102	60
stdJ327A19.1	TCCAGCGATGCAGCTTTAC	ATGGCTAATACCACTTTCCTGC	120	60
stdJ327A19.1n	CAACCAGGAGCTCGAAGCCG	TGCTGCTGAATCTCCAGACT	-	-
stdJ327A19.2	CAAGCCAGAAAGCAATGGAT	GATGACATCCTCAACCAGAGC	121	60
stdJ327A19.3	TGCTGATGAGAAGAGAGCCC	TTTGTCATCCTTCCCTTTGG	120	60
stdJ327A19.3n	CTTGCATCCTTTAACCAAGA	GTCTTTCTGTAAACCATTTC	-	-
stdJ327A19.4	GCCATGCTCTGTTACCTGGT	ACCTGAGCATTCAAATGATGC	124	60
stdJ327A19.4n	TGCAGCTATGGCTGAATATG	ATTTCTTCACTTGTCAAGCC	-	-
stdJ327A19.5	CTACTTGTTCGATCCTTCCAGG	GATGCCCTCTAACATAGGAAGA	170	60
stdJ327A19.5n	ATTTTCCAGAAACTCTTCTT	TTAAGTGAATTTGTATCA	-	-
stdJ327A19.6	AGTTCGTCGAAGAGTCCGAA	AGAGGGCCGCTCTCTAGAAC	188	60
stdJ327A19.6n	AGCCCAGCAAATCTGCCCGC	CGACCGCGAGCGTGAGCGGT	-	-
stdJ327A19.7	GATCGAACAAGTAGGTTC	ATCATCTTGAGAGGTAAG	170	60
stdJ378P9.1	TGAAGGATTTTCAAAGTCTCCA	CATACAAATAGCAACACTGGGC	85	60
stdJ378P9.1n	ATGTTTTGTTGCATTTAG	CATTGTTAATCCTAAATG	-	-
stdJ394H4.1	TTCCAGCAGCCAGTCAAAG	AGGCATGCTGTAGCAGGTG	128	60
stdJ404F18.1	TCAGAGCCCCTACCTCCC	ATTGGCTGTCAATCCATTCC	123	60
stdJ404F18.1n	CACACAGTGGAGGAGTAG	TGGGCCTTCACTATCTGC	-	-
stdJ404F18.2	GTGGACTGCCGCTCTTCTAC	TTAGGAGGCTCTTGTTTCTGAG	132	-
stdJ404F18.2n	CAGCCAAAGAATGCTCCTGT	TGTTCGTTTTCCCAAAGC	-	-
stdJ404F18.3	AGCCAAGAAAGCGAAAATGA	ACGTAAAGCCTTCTGCTAGGG	105	60
stdJ404F18.3n	TGAAGGGGACCTTGATTG	AATAGGAATCCGTCTATG	-	-
stdJ404F18.6	GCATACACGATGCAAGAGGA	GCAGGAGGCACCACTTCTT	135	60

stdJ404F18.6	GCATACACGATGCAAGAGGA	GCAGGAGGCACCACTTCTT	135	60
stdJ452H17.1	ATGCACCTCTGAAACCCTTG	CGACCTCTCTCGGGATATTT	140	60
stdJ452H17.1.1	TTTCCACCACTGGCATTACA	CTCAATAGCCAGGCAGAAGC	159	60
stdJ452H17.1.1n	TGAAACCCTTGCTAAGTA	GGTTGAACTGGAACATAC	-	-
stdJ452H17.1n	AAGTAGAGTATGTTCCAG	AGGCAGAAGCGAATATTG	-	-
stdJ525N14.1	AAAGTTAAAGTCGGCAGGAGC	TCTGGTCGCTGTCCTCAAC	154	60
stdJ525N14.1n	GCCACCTATGGGAAGGAGAC	AACTCCGGCGCCGCCGCCAT	-	-
stdJ525N14.2	TGCCATACACTGGCACTGAT	AAAGAAAGAGCTGGCATCCA	161	60
stdJ525N14.2n	CAACTGGCACACCTCGTTGG	GGCATCCATAGTCGTGGAAG	-	-
stdJ525N14.3	CTTTGGCTTCAGCGCTTC	AACGTCGCTCCAGTCTGG	120	-
stdJ525N14.4	AATGAGCAACGTGGCCAT	GCAACAGAGAAGAGCTGATGG	124	60
stdJ525N14.4n	GACTCTTCTGTGATGTTACC	AGTAGGTACTAGAAGCTGAA	-	-
stdJ525N14.5	AAAAGAACCTCCAGTAGGGACC	TCAAACTCAGTACTGCCATCTG	128	60
stdJ525N14.5n	ACTACATCTACTCAGAACAA	TTACATTTTGTTTAAAAATT	-	-
stdJ525N14.6	AAAGAAGGGGCAGAATCG	TGCTTCCCGGCGCCGCCG	101	60
stdJ525N14.7	ACCCAGACTGGAGCGACG	CTCCTCCGGACGCGCGGAAG	98	60
stdJ525N14.10	ATAGCAATGCCAGTGGAACC	GAGAACACCAGTCTCCGCTC	158	60
stdJ555N2.1	CAGCACCTCTACCTCAAGCC	TGGAGAGCTGAACTGTGGTG	154	60
stdJ555N2.2	TATGGGGTCTTTGCTGGAAG	CTGGGCAGCAGTGAGGTCAG	111	60
stdJ555N2.3	ACAAAAGATTTGGAGGGGCT	AAACTGCTTCCATCCCTGC	141	60
stdJ555N2.3	ACAAAAGATTTGGAGGGGCT	AAACTGCTTCCATCCCTGC	141	60
stdJ555N2.4	GGGCTTGTCAGTGAAATCAA	GAAGATGAGTGAGAGCAAAGGG	142	60
stdJ555N2.4	GGGCTTGTCAGTGAAATCAA	GAAGATGAGTGAGAGCAAAGGG	142	60
stdJ562J12.1	ACATGAAGTTGTTCTCGGGG	CCCTAAGGGTTTTCATCAAGC	144	60
stdJ562J12.1	ACATGAAGTTGTTCTCGGGG	CCCTAAGGGTTTTCATCAAGC	144	60
stdJ562J12.1n	CTGGCTGACCCAAGTCAATG	CAGACATATTCCAATCTGGC	-	-
stdJ655L22.1.1	CTATGCTAGGACACATTAG	GAACTCTGCTTTGTCACAG	130	60
stdJ655L22.1.1n	GATTGGTCATGGAAATAG	GCAGCCCAAAGACTCACATC	130	-
stdJ755D9.1	GAAAACACCGGGGTACACTG	AATGCTGCATGAGAGACATG	120	60
stdJ755D9.1n	GCTGTCACAGACGTCCCA	AATCATGGAGTGCAGTAG	-	-
stdJ755D9.2	AGATGTAAACTAGGAGCAGCCG	CACCAGTGTGAAAGTGAAGAGC	168	60

stdJ755D9.2n	CAGAATGCCGTGGTAGTG	GCTGCTGACTGTCCTCAG	-	-
stdJ755D9.3	ATGGCATCCCCTTAGCTTCT	CTGCCACAGGCTCTCCTC	122	60
stdJ755D9.3n	GTCCGCCAATTATGGCAG	TTATTGAGGAGGCTAGGC	-	-
stdJ755D9.4	TGTTTGTCGGACAAGCTCAG	TGCCTCTTCTTCCTGGCTTA	128	60
stdJ808P6.1	ATGATGAGAATCAGGACCGTG	CCTCCACCATTGCTGTAGGT	127	60
stdJ808P6.1n	CTATGATTGGATACGCAG	AGGTCAGATATGGAAATC	-	-
stdJ808P6.2	ATGGAAAGATGCTGCCACTC	CAAATCCAGCAAACACGCTA	158	60
stdJ808P6.2n	CTTGGGGTGGCTTAGTTG	CGCTAGTGAGACAGTTTG	-	-
stdJ808P6.3	CGCCTTATAAGTTGCTGCAG	AGTAGGTATTTCATGGTCAGCC	120	60
stdJ808P6.3n	ACCTGGAACAACATCGTG	TCTTTCCCCACGATGTTG	-	-
stdJ876A24.1	CTGCCTAGCTCGCGGTCCG	GAAGTCGCCCCCCAACAG	98	60
stdJ876A24.1n	TCGTCCCTGGGGTCCCTG	CTCCAGCCATCTTCTCCG	-	-
stdJ876A24.2	ATGCACAATTCGAGGCCTAC	AGAGACTTCAGGGAATGACCC	127	60
stdJ876A24.2n	CTGCAGGAAGAGCTAAAG	ACAAGCAGACATGGATTCG	-	-
stdJ876A24.3	GAGACCTCGTTTTGAGCCTG	TCTTGGTAGATGTCTCTTGGCA	138	60
stdJ876A24.4	ACACATTTTGCCAGCATG	TGAGAATGAATCCTGATG	109	60
stdJ876A24.4n	TACTCAAGACTCTTTCAG	CTGAGAATTCCCATCTTG	-	-
stdJ878I13.1	GAGAAAGTGGGAGCAGCAAG	ATGCTTCTCTTTCCCTCTTGC	181	60
stdJ1139I1.1	CAGCACCTCTACCTCAAGCC	TGGAGAGCTGAACTGTGGTG	154	60
stdJ1139I1.2	AATGATGGACTCTTCCCGC	ACTTCGTAGGGGTTGACGG	183	60
stdJ1139I1.3	GACGCTCAGCCTCAGCCT	GTTCCCCTTCCACAGGGC	148	60
stdJ1139I1.4	CGTAGACGGGGCTTCCCG	ACCGCGTGGCTCGGGCCTG	109	60
stdJ1139I1.8	GTTCAAAGGAACATTGCCAA	ATAAAGAGTACTTCCTTGGGGG	81	60
stdJ1139I1.8	GTTCAAAGGAACATTGCCAA	ATAAAGAGTACTTCCTTGGGGG	81	60
stdJ1152D16.1	GTTAGAGAAATTGCCCATGAGG	CCGCAGATACATAGTCTCCTCA	109	60
stdJ1152D16.1n	GACCGGTCCCGTGTGATTG	ACTTATGGGCATGTTTGGCAG	-	-
stdJ1189B24.1	GGCGATGGTCCAGATAGAAA	TGGAACCATGAATCCTATTGC	81	60

Table 2.7: *STSs used for mouse mapping as described in chapter 5. STSs are named with the prefix 'st' followed by the clone name from which the STSs was derived. The suffix (SP6, T7 etc.) indicates the vector specific primer used to generate the product.*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)	Gene/BAC
stAA386485.1	AGGAAACCGGAAAAAGGAGA	TCAGTTCGACATCAGAACGC	141	60	EST
stAA597301.1	GTGGACTGCCGCTCTTCTAC	TATTGGTTTTCCCAAAGCCT	105	60	EST
stAB023622.1	AGTCCACGCTCATGGATACC	GTGAGTTTCAACCCCACGTT	129	60	Septin6
stAB023622.2	GGCGAAGATTGTCGAACTGT	GCACAGGATGTTGAAGCAGA	111	60	Septin6
stAB023622.3	CTGCATGAGAAATTTGACCG	CTTTCTCTGCTTGAAGGCGT	111	60	Septin6
stAF042491.1	CCTGCTCTACAAGATCGTTCG	GAACACCTTGCCGTTGATG	175	60	Mapr
stAF089812.1	CGGGAATATGAAAAGCGTGT	TGTCTCTGATGCTCCACAGG	237	60	hHr6a
stAF097416.1	GTGGCCATGGTCTCTTTTGT	AGTTCAACAACTTGCCCAGC	133	60	Znf-kaiso
stU27316.1	TTCGTATCCCCAAGGAACAG	GAACTGGGTCCTCTTGTCCA	216	60	Ant2
stbM65I16SP6	AGCAACCACACTTTGGCTG	ACTCTGCCCTCTTGTGGCTA	170	60	bM65I16
stbM65I16T7	TCCCATGAGTTTACATCTGGC	TTTCACGCACCAACATTCAT	129	60	bM65I16
stbM110K19SP6	AGTGCTGCTGTGTAAAGCAGG	TTTTCTGAGGGACTAAAGGGTT	80	65	bM110K19
stbM110K19T7	TTTGATGCCAGCAGAAAGC	ATGCACCCTGCAGAGTTTCT	150	65	bM110K19
stbM167L6SP6	CAGAGTTGGGAGTCAGTGCA	CTTGGCCTCAAATTAACTCTGG	157	65	bM167L6
stbM193O17SP6	GAATGCTTCATTGGAGGGAA	TGTGACATTTGTCTTACAGCCA	167	60	bM193O17
stbM193O17T7	TGGAACATGGGTCTATCAAGC	GTCAGCCTGTACCCACCATT	120	60	bM193O17
stbM260F21T7	TCCAAACCCTCTGAACCAAC	CACTAAGCAATGGGCCTGAT	266	60	bM260F21
stbM279D19SP6	ATGTCTCTCTGAGGGGCAGA	TGCACTGACTCCCAACTCTG	163	60	bM279D19
stbM286I5SP6	ACTGAGAGGCTAGGAAGCACC	GAAGAATCTTGACTGGGAGGG	125	65	bM286I5
stbM286I5T7	GACATCATGCCCAGGAGG	TTATGCTGCTGCATGACACA	145	65	bM286I5
stbM302H8T7	CAGAGCTTTGCTGTGCAGAG	CTAGCACTGCAAAGGCAGC	135	65	bM302H8
stbM322E15T7	TAGTCTGTAGGGAGGCTCAAGG	CCTCAGGACCACTCTTGAGC	169	65	bM302H8
stbM343N5SP6	GAGACCATATCCAGTGGCTAGG	GCAGTCTCTGGATGGCTTTC	152	65	bM343N5
stbM461E19T7	GTCATGCATGGTAGTTGCATG	CAGTAAGCTGAACTTAGGTGCA	160	65	bM461E19

Table 2.8: *STSs for conserved sequence analysis in human and mouse described in chapter 5*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)
stdJ404F18.6	GCATACACGATGCAAGAGGA	GCAGGAGGCACCACTTCTT	135	60
stdJ555N2.3	ACAAAAGATTTGGAGGGGCT	AAACTGCTTCCATCCCTGC	141	60
stdJ555N2.4	GGGCTTGTCAGTGAAATCAA	GAAGATGAGTGAGAGCAAAGGG	142	60
stdJ1139I1.7	TATTACTGAGGGAAACAGCTGG	ATGTGAAGCTGTCCAGTTTTTT	89	60
stdJ1139I1.8	GTTCAAAGGAACATTGCCAA	ATAAAGAGTACTTCCTTGGGGG	81	60

Table 2.9: *STSs used for zebrafish mapping as described in chapter 6*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)	Gene
stbG421I3.4	GGCAATATGCATGGATGTGA	CAAGGTCTGAGGCAGGTTTC	100	60	bG421I3.CX.2
stbG421I3.5	GAACCTGGGCAGCAGTATTC	AAAATGCACTCCAGCTCCTG	122	60	dJ525N14.CX.1
stbK38K21.3	ACTTTCAGGATTAAGCGATTCC	TCCAGTTTTCATCCGAATCC	81	65	RPL39
stdJ327A19.10	AGCCAAGAAATTGGACAAAG	ATCTTTAAGTTCGCTATCAG	94	60	UPF3B
stdJ327A19.11	TGGCCATGCTCTGTTACCTG	TCACTTGTCAAGCCAATTTC	99	60	dJ327A19.CX.3
stdJ327A19.12	GGTCACAGACATAGCAGCGC	CGTTCAGATTACAAGCATG	235	60	ZNF183
stdJ327A19.14	CTGTGTCCAGCTCATAGACAGC	CGACTTGAGCAGCGAAGAG	131	65	ZNF183
stdJ327A19.13	GTTAGTGAACCTGTGGATG	GTAACGGGGCAGAGATGTG	107	60	NDUFA1
stdJ404F18.4	AGATCTTCCTGGGTGGTGTG	TGCACAGACACGTTAAAGCC	256	60	ANT2
stdJ404F18.5	GTGGACTGCCGCTCTTCTAC	TATGTTCGTTTTCCCAAAGC	107	60	dJ876A24.CX.1
stdJ525N14.11	TGTGGAAGACCGAAAATTCC	TCAGTTCAACAACTTGCCCA	101	60	KAISO
stdJ555N2.2	TATGGGGTCTTTGCTGGAAG	CTGGGCAGCAGTGAGGTCAG	111	60	HPR6.6
stdJ876A24.2	ATGCACAATTCGAGGCCTAC	AGAGACTTCAGGGAATGACCC	127	60	SEPTIN2
stdJ876A24.11	ACAGATTGCAAGGCAACG	TGAATTCTCCAGGGGAAG	368	60	SEPTIN2
stdJ876A24.16	TTCTCCAAATGGCTGAAGGT	ACATGTTGAGAGGTGGCGTT	101	60	dJ876A24.CX.3
stdJ876A24.17	CTCTGTTGGATGAACCCAATC	CAATCACGCCAGCTTTGTTC	124	60	UBE2A
stdJ1139I1.6	ACAGACCTCATCAAAACCCG	AAGGAACCCTTCCTGTTGGT	105	60	dJ1139I1.CX.1
stwz3779.1	ACACTGCTGCTTTGCTGAGA	GGAGGCAGTGAAGAAGTTGC	156	60	dJ876A24.CX.3
stwz8217.1	TAGCTTGGCTCGTTCTTGGT	GTGTCGTGATTTGTGCTCGT	247	60	dJ327A19.CX.3

2.10 World Wide Web addresses

Baylor College of Medicine Search Launcher	http://dot.imgen.bcm.tmc.edu:9331/
Baylor College of Medicine Sequencing Center	http://www.hgscbcm.tmc.edu/
British Columbia Genome Sequence Centre	http://www.bcgsc.bc.ca/
CHLC	http://lpg.nci.nih.gov/CHLC/
DOTTER	http://www.cgr.ki.se/cgr/groups/sonhammer/Dotter.html
EMBL	http://www.ebi.ac.uk/
GDB	http://gdbwww.gdb.org/
Généthon	http://www.genethon.fr/genethon_en.html
Genome Sequencing Center, St Louis	http://www.ibc.wustl.edu/cgm/jcgm.html
Genome Sequencing Center, Jena	http://genome.imb-jena.de/
Genome Sequencing Center, Naples	http://hpced.area.na.cnr.it/grsl/
INTERPRO	http://www.ebi.ac.uk/interpro/scan.html
MPIMG, Berlin (X sequencing)	http://www.mpimg-berlin-dahlem.mpg.de/~xteam/
National Centre for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
OMIM	http://www3.ncbi.nlm.nih.gov/Omim/
PIPMAKER	http://bio.cse.psu.edu/cgi-bin/pipmaker
RepeatMasker	http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html
The Institute for Genome Research	http://www.tigr.org/
The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/
TRANSFAC	http://transfac.gbf.de/TRANSFAC/
Washington University Center for Genetics in Medicine (CGM)	http://www.ibc.wustl.edu/cgm/
Whitehead Institute	http://www-genome.wi.mit.edu/
X Chromosome Mapping Project at the Sanger Institute	http://www.sanger.ac.uk/HGP/ChrX/
VISTA	http://www-gsd.lbl.gov/vista
Zebrafish RH mapping	http://www.genetics.wustl.edu/fish_lab/cgi-bin/human_int_map.cgi

Methods

2.11 Isolation of bacterial clone DNA

2.11.1 Miniprep of cosmid, PAC and BAC DNA

1. Ten ml of 2 X TY containing 30 µg/ml of appropriate antibiotic (kanamycin for cosmids and PACs, chloramphenicol for BACs) were inoculated with a scraping from the frozen glycerol stock of the chosen bacterial clone and incubated overnight at 37°C with shaking.
2. The cells were collected by centrifugation at 4,000 rpm for 10 minutes at room temperature in a Beckman J6-MC, resuspended in 200 µl of GTE in a 1.5 ml eppendorf tube, and left on ice for 5 minutes.
3. 400 µl of freshly prepared 0.2 M NaOH/1% SDS were added to the cells, mixed by gentle inversion, and the sample left on ice for another 5 minutes.
4. 300 µl of 3 M K^+/5 M Ac^- (pH 4.8) were added, mixed by gentle inversion and left on ice for 10 minutes. The sample was centrifuged for 10 minutes at 14,000 rpm in an Eppendorf microfuge.
5. The supernatant was transferred to a fresh tube and mixed with 600 µl of cold isopropanol and left on ice for at least 10 minutes. The tube was subjected to centrifugation for 15 minutes at 14,000 rpm in an Eppendorf microfuge at 4°C to pellet the DNA, the supernatant removed and the pellet resuspended in 200 µl of $T_{0.1}E$.
6. 200 µl of 50:50 (v/v) phenol/chloroform were added to the sample, which was vortexed and briefly centrifuged. 20 µl of 3 M sodium acetate (pH 5.2) and 200 µl of isopropanol were added to the aqueous layer, and the sample placed at -20°C for at least 10 minutes. The tube was subjected to centrifugation at 14,000 rpm in an Eppendorf microfuge for 15 minutes at 4°C to pellet the DNA. The pellet was washed with 70% ethanol and resuspended in 50 µl of $T_{0.1}E$.
7. 1 µl of 10 mg/ml RNase was added and the sample incubated at 37°C for 1 hour prior to storage at -20°C.

2.11.2 Microprep of cosmid, PAC and BAC DNA for restriction digest fingerprinting

1. 500 µl of 2 X TY containing appropriate antibiotic (see Section 2.11.1) were added to a 96 well deep-well microtitre plate (COSTAR).

2. Each well was inoculated from a glycerol stock with either a 96-well inoculating tool, or a sterile cocktail stick. A plate sealer (Dyntax) was placed on top of the plate to seal the wells, and the culture grown for 18 hours at 37°C with gentle shaking.
3. For each well, 250 µl of the overnight growth were transferred to a clean microtitre plate. The cells were collected by centrifugation (Sorvall RT7, Du Pont Company Sorvall, Delaware US) at 1550 g for 4 minutes.
4. For each well, the supernatant was removed and the pellet resuspended in 25 µl of GTE, by vortexing gently (a cocktail stick was used for resuspending pellets still attached to the plate).
5. 25 µl of GTE were added to each well and gently mixed. 25 µl of freshly prepared 0.2 M NaOH/1% SDS were added, mixed and left to stand for 5 minutes at RT.
6. 25 µl of 3 M K^+/5 M Ac^- (pH5.0) were added, mixed and left at RT for 5 minutes. A plate sealer was placed on top of the plate and the plate was vortexed gently for 10 seconds.
7. A microtitre plate containing 100 µl of isopropanol was taped to the bottom of 2 µm filter-bottomed plate (Millipore cat. no. MAGVN2250). The total well volume of the sample was transferred to the filter-bottomed plate and the sample was filtered by centrifugation at 1550 g for 2 minutes at 20°C.
8. The filter-bottomed plate was removed and the microtitre plate was left at RT for 30 minutes, before being centrifuged at 1500g for 20 minutes at 20°C.
9. The supernatant was removed and the DNA was dried by inverting the plate and placing it on clean tissue paper, ensuring no disruption of the pellet.
10. 100 µl of 70% ethanol were added to the dried DNA, mixed gently, and DNA precipitated by centrifugation at 1500g for 10 minutes at 20°C. For restriction digest fingerprinting the wash was repeated. The supernatant was removed and the DNA dried as before.
11. 5 µl of freshly prepared $T_{0.1}E$ / 1 µg/ml RNase were added and mixed gently to resuspend the DNA. Samples were stored at -20°C.

2.12 Bacterial clone fingerprinting

2.12.1 Radioactive fingerprinting

1. For each 96-well microtitre plate of sample DNA, a premix containing 1x NEB2 buffer (New England Biolabs), 0.72 U *Hind* III, 1.3 U *Sau*3AI, 0.4 U Reverse Transcriptase, 0.07 µl [alpha-^{32}P]dATP (3000Ci/mmol), 0.04 µl 10 mM ddG was prepared in a 1.5 ml microfuge tube.
2. 2 µl of premix were added to the sample DNA using a Hamilton repeat dispenser. The reaction was mixed by gentle agitation and the plate was spun at 150 *g* for 10 seconds (Sorvall RT7, Du Pont Company Sorvall, Delaware US).
3. The reaction was incubated for 1 hour at 37°C, and the reaction stopped by the addition of 2 µl formamide dye.
4. The sample DNA was denatured at 80°C for 10 minutes and loaded in groups of 6, leaving the first well and every subsequent seventh well of a 4% polyacrylamide gel empty (see Section 2.14.2). Marker DNA (see Section 2.13.1) was denatured by boiling for 5 minutes and 2 µl were loaded in the first well and every seventh well. Fragments were resolved by running the gel at 74 W for 1.5 hours (or until the bromophenol blue dye front reached the bottom of the gel).
5. Following electrophoresis, the back plate was removed and the gel was fixed in a 10 % glacial acetic acid solution for 10 minutes, then washed in water for 25 minutes. The gel was dried onto the front plate by incubation at 80°C for 45 minutes in an oven. Autoradiography was for 72 hours at RT.
6. The autoradiograph was scanned using a flat bed scanner (Amersham) and the digitised version imported to IMAGE.

2.12.2 Fluorescent fingerprinting

1. For one 96-well microtitre plate of sample DNAs, three digest premixes were prepared, one for each fluorescent label, in three 1.5 ml microfuge tubes labelled TET, HEX and NED. Each premix contained 25.5 µl $T_{0.1}E$, 24.5 µl NEB2 buffer, 5.0 µl *Hind* III (20 U/µL), 8.0 µl Taq FS, (32 U/µl) and 3.0 µl *Sau*3AI (30 U/µl), 4.0 µl of the appropriate ddA-dye. Each premix was mixed prior being aliquoted.
2. 2 µl of the TET premix were added to wells A1-H4 of the microtitre plate containing sample DNAs using a Hamilton repeat dispenser. Similarly, 2 µl of the HEX premix were added to wells A5-H8 and 2 µl of the NED premix were added to wells A9-H12.

The plate was covered with a plate sealer, the reaction mixed by gentle agitation on a vortex. In order to ensure the sample was in the bottom of the wells the plate was centrifuged at 150 *g* for 10 seconds (Sorvall RT7, Du Pont Company Sorvall, Delaware US).

3. The reaction was incubated for 1 hour at 37°C.
4. To precipitate the DNA, 7 µl 0.3*M* sodium acetate and 40 µl 96% ethanol were added to each well. For multiplexing the samples, rows 5 and 9 were added to row 1, rows 6 and 10 were added to row 2, rows 7 and 11 were added to row 3, and rows 8 and 12 were added to row 4 respectively, using a multichannel pipette.
5. The samples were incubated at RT for 30 minutes in the dark.
6. The plate was subjected to centrifugation at 1550 *g* for 20 minutes at 20°C to pellet the DNA.
7. The supernatants were discarded and the pellets dried by tapping the plate face down onto tissue paper.
8. The pellets were washed by adding 100 µl of 70% ethanol to each well, mixed gently tapping the plate, and the plate was subjected to centrifugation at 1550 *g* for 10 minutes at 20°C.
9. The supernatants were discarded and the pellet dried as above.
10. The DNAs were resuspended in 5 µL $T_{0.1}E$.
11. Prior to loading, 2 µl of the marker DNA (see Section 2.13.2) were added to each sample using a Hamilton repeat dispenser. The samples were denatured for 10 minutes at 80°C. 1.25 µl of each sample were loaded on a 5% denaturing acrylamide gel and resolved on a ABI377 Automated DNA sequencer using a 0.2 mm, 12cm, well-to-read 4.5% denaturing polyacrylamide gel (prepared by Sanger Institute Gel Production team). Data were collected using the ABI Prism Collection Software v1.1.
12. After data collection, the gel image was transferred to a UNIX workstation for entry into IMAGE.

2.12.3 Hind III fingerprinting

1. For one 96-well microtitre plate of sample DNA, a premix containing 231 µl H_20, 99 µl buffer B, 55 µl *Hind* III, was prepared in a 1.5 ml microfuge tube, and mixed using a vortex. 4 µl of the premix were added to each well of a 96-well microtitre plate containing previously prepared DNA (see Section 2.11.2), and the plate covered with a plate sealer (Dynex).

2. The reaction was mixed gently on a vortex and incubated at 37 °C for 2 hours.
3. The reaction was terminated by the addition of 2 µl of buffer II and either loaded straight away or stored at 4 °C.
4. 0.8 µl of the marker (see Section 2.13.3) were added to the first well and then every sixth well of a freshly prepared 1% agarose/1x TAE gel (see Section 2.14.1 for preparation). 1 µl of each sample was loaded (i.e. wells 2-5, 7-10 *etc.*) between the marker lanes. Fragments were resolved by running the gel at 4°C in a cold room for 15 hours at 90 volts.
5. Following electrophoresis, the gel was cut down so the length was 19-20 cm and stained in Vista Green (mix 5 ml 1*M* Tris HCL, 0.5 ml 0.1*M* EDTA, 50 µl Vista Green, make up to 500 ml with H_20) for 30-45 minutes on a shaker. The gel was washed with H_20 to remove excessive stain.
6. The gels were scanned on a FluorImager SI. The parameters were set to 530 nm for emission filter, the pixel size was 100 microns, detection sensitivity was normal, digital resolution was at 16 bits, dye was single label, excitation filter was 488 nm, Em filter 1530 nm and PMT voltage was 800.
7. The gel image was transferred to a UNIX workstation for entry into IMAGE.

2.13 Marker preparation

2.13.1 For radioactive fingerprinting

1. 68 µl of T0.1E, 10 µl of NEB2 buffer, 6.6 µl of Sau3AI (50 U/µl), 4 µl dGTP, 5 µl ddTTP, 0.07µl [alpha-^{35}S]dATP (3000Ci/mmol), 2 µl Reverse Transcriptase were added to a 1.5 ml microfuge tube and incubated at 37°C for 1 hour.
2. The reaction was stopped by the addition of 106 µl of 1:15 dilution of formamide dye.
3. The marker was stored at -20°C.

2.13.2 For fluorescent fingerprinting

1. 70 µl $T_{0.1}E$, 10 µL NEB2, 6 µl lambda DNA (500 ng/µl), 6 µl *Bsa*J1 (2.5 U/µl), 4 µl TaqFS (32 U/µl), 4 µl ddC-ROX were added to a 1.5 ml microfuge tube and incubated for 1 hour at 60°C.

2. 100 µl 0.3 *M* sodium acetate and 400 µl 96% ethanol were added to the reaction mix and incubated at room temperature in the dark for 15 minutes, then at -20°C for 20 minutes. The tube was subjected to centrifugation in a bench top centrifuge at maximum for 20 minutes to pellet the DNA.
3. The supernatant was discarded and the DNA pellet dried by tapping the tube gently onto tissue paper. The pellet was washed by adding 200 µl 70% ethanol and spun in a bench top centrifuge at maximum for 5 minutes, the supernatant discarded and the pellet dried as described in step 2.
4. The DNA was resuspended in 120 µl $T_{0.1}E$ and 120 µl blue dextran formamide dye.
5. The marker was stored at -20°C.

2.13.3 For Hind III fingerprinting

1. 19.2 µl $T_{0.1}E$, 1.5 µl Analytical Marker DNA wide range, 0.2 µl Molecular Weight marker V and 4.2 µL 6x loading dye were added to a 1.5 ml microfuge tube.
2. The marker was stored at -20°C.

2.14 Gel preparation and electrophoresis

2.14.1 Agarose gel preparation and electrophoresis

1. Agarose gels were prepared in 1x TBE (or 1x TAE, for *Hin*d III fingerprinting) containing 250 ng/µl ethidium bromide and the appropriate percentage of agarose according to the size of fragments being separated: 2.5 % agarose gels were used for electrophoresis of fragments below 1 kb; 1.0% agarose gels were used for analysis of larger fragments. Electrophoresis was performed at 50 - 90 V for 15 - 45 minutes depending on the separation required.

2.14.2 Polyacrylamide gel preparation for radioactive fingerprinting

1. 42.0 g urea were dissolved in 10 ml 10x TBE and 35 ml ddH_20 by warming to 37°C, and stirring.
2. A large glass plate (back plate – Gibco BRL) was washed on both sides and one side was treated with 2 % dimethyldichlorosilane, and left to dry.

3. A small glass plate (front plate – Gibco BRL) was washed on both sides with detergent and water and one side was treated with freshly prepared ethanol/acetic acid/WAKKER solution (3 ml 96 % ethanol, 50 µl 10 % Acetic acid, 5 µl Walker soln) and left to dry.
4. The front and back plates were taped together along three edges (treated sides facing inwards) separated by 4mm spacers.
5. 10 ml 40% acrylamide, 800 µl 10 % ammonium persulphate and 80 µl TEMED (KODAK) were added to the dissolved urea solution, mixed and poured in between the glass plates using a 50 ml syringe. A 4mm, 60 well comb (IBI) was placed in the top of the gel (the edge not taped) and the glass plates clamped with bulldog clips. The gel was left to set for up to 3 hours.

2.15 Applications using the polymerase chain reaction

2.15.1 Primer design

Primers were designed manually using the following guidelines:

1. As far as possible, sequences chosen were 18 - 25 bp in length, beginning and ending with a C or G.
2. Sequences were chosen to avoid areas of simple sequence showing non-representative use of the bases and obvious repetitive sequence i.e., runs of single nucleotide (e.g. TTTT) or double nucleotide (CGCGC) motifs.
3. Sequences were chosen to exclude palindromes which will form inhibitory secondary structure, especially at the 3' ends (e.g. GACGTC).
4. As far as possible, sequences were chosen with a GC content of at least 50%.
5. Sequences were chosen to avoid complementarity between pairs of primers, especially at the 3' end, which could result in primers annealing to each other and forming primer dimers.
6. If possible, sequences were chosen which would generate products of at least 100 bp in length.

215.2 Oligonucleotide preparation

All oligonucleotides used were synthesised in house by David Fraser or supplied as working dilutions from Genset. The concentration of the primer in ng/µl was determined by measuring the absorbance at 260 nm (Abs_{260}) and multiplying this by 33 and any necessary dilution factor.

2.15.3 Amplification of DNA by PCR

1. 1-3 ng/µl of genomic DNA were amplified in a reaction volume of 15 to 50 µl as required. Reactions contained approximately 1.3 µM of each oligonucleotide primer, 67 mM Tris-HCl (pH 8.8), 16.6 mM $(NH_4)_2S0_4$, 6.7 mM $MgCl_2$, 0.5 mM of each deoxyribonucleoside triphosphate (dATP, dCTP, dGTP, dTTP), 1.5 U of Amplitaq™ (Cetus Inc.). 10 mM β-mercaptoethanol and 170 µg/ml of BSA (Sigma Chemical Co., A-4628) were added to the reactions from freshly made stock solutions as the reactions were set up.

2. Unless specified otherwise, cycling conditions were as follows: all reactions were preceeded by an initial denaturing step of 5 minutes at 94°C, followed by 35 cycles of: 93°C for 30 seconds, [primer-specific annealing temperature] for 30 seconds, and 72°C for 30 seconds; followed by a final extension step of 5 minutes at 72°C. Primer-specific annealing temperatures are given for each primer pair in the text or in Tables 2.3 – 2.9.

3. PCR products were separated on 2.5% agarose minigels as described in Section 2.14.1 and visualised by ethidium bromide staining.

2.15.4 Colony PCR of STSs from bacterial clones

1. Colony PCR on bacterial clones was performed by touching a sterile toothpick onto the surface of a colony and stirring this into 200 µl of $T_{0.1}E$, and using 5 µl of the resulting suspension in a 15 µl final volume PCR.

2. PCR products were separated on 2.5% agarose minigels as described in Section 2.14.1 and visualised by ethidium bromide staining.

2.16 Radiolabelling of DNA probes

2.16.1 Random hexamer labelling

adapted from Feinberg, A. P., *et al.*, 1983

1. Approximately 10 ng of DNA were boiled in a total volume of 13.5 µl of $T_{0.1}E$ for 5 minutes and snap chilled on ice-water.
2. Following the addition of 5 µl of OLB3, 1 µl of 10 mg/ml BSA, 2.5 U of Klenow enzyme (Boehringer Mannheim, sequencing grade) and 1 to 5 µl of [α-^{32}P]-dCTP (3,000 Ci/mmol; 10 Ci/ml) and $T_{0.1}E$ to a total volume of 25 µl, the reactions were mixed and left at room temperature for a minimum of 3 hours and up to overnight.
3. All probes were boiled for 5 minutes and snap-chilled on ice-water prior to use.

2.16.2 Radiolabelling of PCR products by PCR

PCR products were radiolabelled essentially as described in Bentley *et al.* (1992).

1. 5 - 10 µl of PCR product were separated on a 2.5% agarose minigel and visualised by ethidium bromide staining.
2. The gel was rinsed in deionised water to remove excess buffer. The desired band was excised from the gel and placed in 100 µl of $T_{0.1}E$ at 4°C overnight.
3. 2 µl of the $T_{0.1}E$ were used as template in the PCR-labelling reaction containing 40 ng of each primer, 1 µl of 10x PCR buffer, 0.5 µl of [α-^{32}P]-dCTP (3,000 Ci/mmol), 0.5 U of *Taq* polymerase (Cetus) and 0.375 mM each of dATP, dTTP and dGTP. Reactions were performed in a 0.5 ml microfuge tube and overlaid with mineral oil (Sigma) in a DNA thermal cycler (Perkin Elmer, USA).
4. PCR cycling conditions were as follows: 94°C for 5 minutes; followed by 20 cycles of: 93°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds; followed by 72°C for 5 minutes.
5. Probes were pre-reassociated (as described in Section 2.16.3) prior to use if necessary. All probes were boiled for 5 minutes and snap-chilled on ice prior to use.

2.16.3 Pre-reassociation of radiolabelled probes

1. Radiolabelled probe was mixed with 125 µl of 20x SSC and 250 µl of the sheared 10 mg/ml human placental DNA (Sigma) in a final volume of 500 µl.
2. The mix was boiled for 5 minutes, snap-chilled in ice-water, then added directly to the hybridisation reaction.

2.17 Hybridisation of radiolabelled DNA probes

2.17.1 Hybridisation of DNA probes derived from whole cosmids

1. Filters were prehybridised flat in sandwich boxes in 50 ml of hybridisation buffer at 65°C with gentle shaking.
2. Radiolabelled probe was denatured by boiling for 5 minutes, added to the sandwich box, and hybridised to the filters at 65°C for 18 hours.
3. Filters were washed twice at RT in 2x SSC for 5 minutes, twice at 65°C in 0.5 x SSC, 0.1% Sarkosyl for 30 minutes, twice at 65°C in 0.2x SSC, 0.1% Sarkosyl for 30 minutes, and twice at 65°C in 0.1x Sarkosyl, 0.1% Sarkosyl for 30 minutes. All washes were carried out with gentle shaking. Filters were rinsed at RT in 2x SSC to remove Sarkosyl.
4. Excess liquid was removed from the filters by laying them briefly on Whatman 3MM paper. Filters were then wrapped in Saran Wrap (Dow Chemical Co.) and exposed to autoradiograph film under the appropriate conditions.

2.17.2 Hybridisation of DNA probes derived from STSs

1. Filters were prehybridised tightly rolled in 15 ml Sterilin tubes or flat in sandwich boxes for 3 hours in 10-25 ml of hybridisation buffer at 65°C with gentle shaking.
2. Radiolabelled probe was added and hybridised to the filters as described in step 2 of Section 2.17.1
3. Filters were washed twice at RT in 2x SSC for 5 minutes, twice at 65°C in 0.5 x SSC, 0.1% SDS for 30 minutes. Filters were rinsed at RT in 0.2x SSC prior to draining the excess liquid, wrapping in Saran wrap (Dow Chemical Co.) and exposing to autoradiograph film.

2.17.3 Hybridisation of DNA probes to gridded zebrafish library

1. Filters were prehybridised in sandwich boxes for 3 hours in 10-25 ml of hybridisation buffer at 50°C with gentle shaking.
2. Radiolabelled probe was denatured by boiling for 5 minutes, added to the sandwich box, and hybridised to the filters as described in step 2 of Section 2.17.1
3. Filters were washed twice at room temperature in 6x SSC for 5 minutes, twice at 50°C in 6 x SSC, 0.1% SDS for 30 minutes, followed by similar sequential washes reducing the concentration of SSC to 4x, 2x, and 1x. Washes were stopped when the amount of the signal remaining on the filters reached less than 5 cpm (tested using a Gieger counter). Filters were given two final rinses in 1 x SSC prior to draining the excess liquid, wrapping in Saran Wrap (Dow Chemical Co.) and exposing to X-ray film.

2.17.4 Stripping radiolabelled probes from hybridisation filters

1. Filters were washed in 0.4 M NaOH for 30 minutes at 42°C followed by 30 minutes in 0.2 M Tris-HCl (pH 7.4), 0.1x SSC, and 0.1% w/v SDS at 42°C with gentle shaking. Successful removal of radiolabelled probe was assessed by autoradiography.

2.18 Restriction endonuclease digestion

2.18.1 Restriction endonuclease digestion of cosmid DNA

1. 4 µl (approximately 150 ng) of prepared cosmid DNA (described in Section 2.11.1) were digested with *Hin*d III using commercial buffers according to manufacturers' instructions in a final volume of 10 µl.
2. 5 µl of each digest were checked for complete digestion by electrophoresis on a 1% agarose minigel and visualised by ethidium bromide staining. The remaining 5 µl were either used immediately for whole cosmid hybridisation (see Section 2.17.1) or stored at –20°C.

2.18.2 Restriction endonuclease digestion of PAC or BAC DNA

1. 5 µl (approximately 200ng) of prepared PAC or cosmid DNA (described in Section 2.11.1) were digested with *Rsa* I using commercial buffers supplied according to manufacturers' instructions, but with the addition of 1 mM spermidine, in a final volume of 20 µl.
2. 10 µl of each digest were checked for complete digestion by electrophoresis on a 1% agarose minigel and visualised by ethidium bromide staining. The remaining 10 µl were either used immediately for vectorette library construction (see Section 2.19) or stored at –20°C.

2.19 Generation of vectorette library of PACs and BACs

Adapted from vectorette PCR of YACs (Riley, J., *et al.*, 1990)

1. *Rsa* I-digested PAC or BAC DNA (see Section 2.11.1) was precipitated by adding 40 µl of ddH_20, 5 µl of 3M sodium acetate and 100 µl of cold (–20°C) 96% ethanol, and leaving for 1 hour at –20°C. The microfuge tube was subjected to centrifugation in a bench top microfuge for 15 minutes at 14,000 rpm to pellet the DNA. The supernatant was discarded and the DNA pellet washed in 70% ethanol, and air dried.
2. The DNA pellet was resuspended in 100 µl freshly made ligation buffer. 10 µl of annealed vectorette bubbles (BPBI and BPHII, 1pm/µl), 1.1 µl rATP and 2.5 units T4 DNA ligase were added to the sample which was incubated at 37°C for 1 hour.
3. The vectorette library was diluted with 400 µl with $T_{0.1}E$, and was stored at –20°C.

2.20 Rescue of clone ends by PCR amplification of vectorette libraries

Adapted from vectorette PCR of YACs (Riley, J., *et al.*, 1990)

1. PCR was performed using 1 µl of vectorette library as template. The primers used were either PACT2 (T7 end) or PACS2 (SP6 end) in conjunction with 224. Reactions were carried in standard buffer conditions described in Section 2.15.3.
2. PCR was performed in a DNA thermocycler (MJ). Cycling conditions were as follows: an initial denaturation step of 5 minutes at 94°C, followed by 35 cycles of: 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 3 minutes; followed by a final step of 10 minutes at 72°C.

3. 5 µl of PCR product were separated by electrophoresis through 2.5% or 1% agarose gels and visualised by ethidium bromide staining. Bands were excised from the gel and placed in 100 µl of $T_{0.1}E$ and stored at 4°C until required. Products for direct sequencing were gel purified using either GeneClean™ or Qiagen™.

2.21 Preparation of colony grids

1. Clones were spotted onto Hybond N filters using sterile cocktail sticks. The filters were incubated at 37°C overnight.
2. Filters were transferred sequentially onto Whatman 3 MM paper soaked in the following solutions for the times given: 10% SDS (4 minutes), denaturation solution (5 minutes), neutralisation solution, (5minutes), 2x SSC/0.1% SDS (5 minutes), 2x SSC (5 minutes).
3. Filters were air dried on Whatman 3 MM paper. Prior to use in hybridisations the DNA was UV-cross-linked colony side down for 2 minutes on a transilluminator (320 nm).

2.22 Clone library screening

2.22.1 Clone library screening with STSs

1. Pools of bacterial clone DNA, each containing DNA from 3072 clones from 8 x 384-well-microtitre plates, were prepared by E. Sotheran and D. Pearson from the RPCI-1, 3, 4, 5, 6, 11 and 13 libraries.
2. The DNA pools were arranged in microtitre plates for screening. In the primary screen, 5 µl of each pool were used as template in a 15 µl final volume PCR using buffer and PCR conditions as described in Section 2.15.3. 50 ng of genomic DNA and $T_{0.1}E$ were included as positive and negative controls respectively. 10 µl of the PCR products were loaded on 20 cm x 20 cm 2.5% agarose horizontal slab gels using an 8-way multi-channel pipetting device, separated by electrophoresis and visualised by ethidium bromide staining.
3. In the secondary screen, the PCR product was radiolabelled as described in Section 2.16.2 and hybridised to individual filters of gridded arrays of bacterial clones representing the pools in which a positive signal had been observed, as described in Section 2.17.2

4. Positive clones were picked from the library and streaked to single colonies on LB agar plates with appropriate antibiotic and grown overnight at 37°C. Single colonies were confirmed by colony PCR as described in Section 2.15.3

2.22.2 cDNA library screening by PCR

The strategy used to screen the cDNA libraries by PCR is illustrated in Figure 2.1.

1. Twenty different cDNA libraries were subdivided into 25 subpools of 20,000 clones, which were then combined to produce 5 superpools of 100,000 clones by J. Bye and S. Rhodes. Details of the cDNA libraries are given in Table 2.2.
2. Aliquots of the superpools of each library were arranged in a microtitre plate to facilitate subsequent manipulations and gel-loading post PCR with a multi-channel pipetting device.
3. In the primary screen, 5 µl of each superpool were used as template in a 15 µl final volume PCR using buffer and PCR conditions as described in Section 2.15.3.
4. PCR products were loaded on 20 cm x 20 cm 2.5% agarose horizontal slab gels using an 8-way multi-channel pipetting device, separated by electrophoresis and visualised by ethidium bromide staining.
5. In the secondary screen, 5 µl of each of the 5 subpools of 20,000 clones corresponding to the superpool that were positive in the first round, were screened by PCR with the same primer pair as used in step 2. PCR products were separated by electrophoresis through 2.5% agarose minigels and visualised by ethidium bromide staining.

2.22.3 Single-sided specificity PCR (SSPCR) on cDNA libraries

The principle of SSPCR (Huang, S.-H., *et al.*, 1993) is illustrated in Figure 2.1.

1. SSPCR was performed on the subpools of the cDNA libraries, each containing 20,000 clones. Prior to their use in PCR, the subpools were diluted 1:10 in $T_{0.1}E$ and boiled. Dilutions were stored at -20°C until required. On removing from -20°C, tubes were centrifuged briefly in a microfuge to settle the contents and then mixed carefully when thawed.
2. In the first round, PCR was performed using 1 µl of the diluted subpools as template in a 15 µl final volume using buffer conditions as described in Section 2.15.3. The primer combinations used are given in Table 2.10.
3. PCR was performed in microtitre plates in a DNA thermocycler (Omnigene) using hot-start. Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95°C, followed by 25 cycles of: 94°C for 30 seconds, 60 °C for 30 seconds, and 72°C for 3 minutes; followed by a final step of 10 minutes at 72°C.
4. For the second round of PCR, products from the first round were diluted 1 in 50 and 1 in 500 in $T_{0.1}E$. 5 µl of each dilution was used as template in 15 µl final volume PCR using buffer conditions as described in Section 2.15.3. Cycling conditions were as described in step 3. The primer combinations used are given in Table 2.10.

Table 2.10: *Primer combinations used in SSPCR**

First round SSPCR	Second round SSPCR
Specific primer A and FP vector primer	Specific primer B and FP vector primer
Specific primer A and RP vector primer	Specific primer B and RP vector primer
Specific primer C and FP vector primer	Specific primer D and FP vector primer
Specific primer C and RP vector primer	Specific primer D and RP vector primer

*Primer sequences are given in Table 2.3

5. 5 µl of the second-round PCR products were separated by electrophoresis through either 1% or 2.5% agarose minigels depending on product size and visualised by ethidium bromide staining. Products were gel purified using the Qiaquick gel extraction kit (Qiagen™) prior to sequencing directly.

Figure 2.1: *(see over) Strategy for SSPCR on cDNA libraries (a) Superpools representing 100,000 clones (A-E) were screened by PCR and positive super pools recorded (e.g. D, shown in red).The five pools(containing 20,000 clones) that were combined to form the positive superpool (e.g. 16-20) were screened and positive pools recorded (e.g.16, shown in red).(b)Rescue of the insert of the cDNA of interest by SSPCR. cDNA is shown as white rectangle representing insert, vector shown as grey rectangles. Original position of primers for pool screening is indicated by a red rectangle. A combination of vector primers (e.g.FP and RP) and insert specific primers (e.g A and D) were used in the first round of the SSPCR. A second round of SSPCR was carried out with nested sequence specific primers (B and C) to generate two products (e.g. C+FP, B+RP) representing the entire insert of the cDNA of interest.*

(a)

Superpool Screen – 5x 100,000 clones
for each library
A
B
C
D
E
Pool Screen – 5x 20,000 clones
for each library
16
17
18
19
20

(b

VECTOR
cDNA
VECTOR
Forward
Primer
Reverse
Primer
specific
primer A
specific
primer B
specific
primer C
specific
primer D
C+FP
B+RP

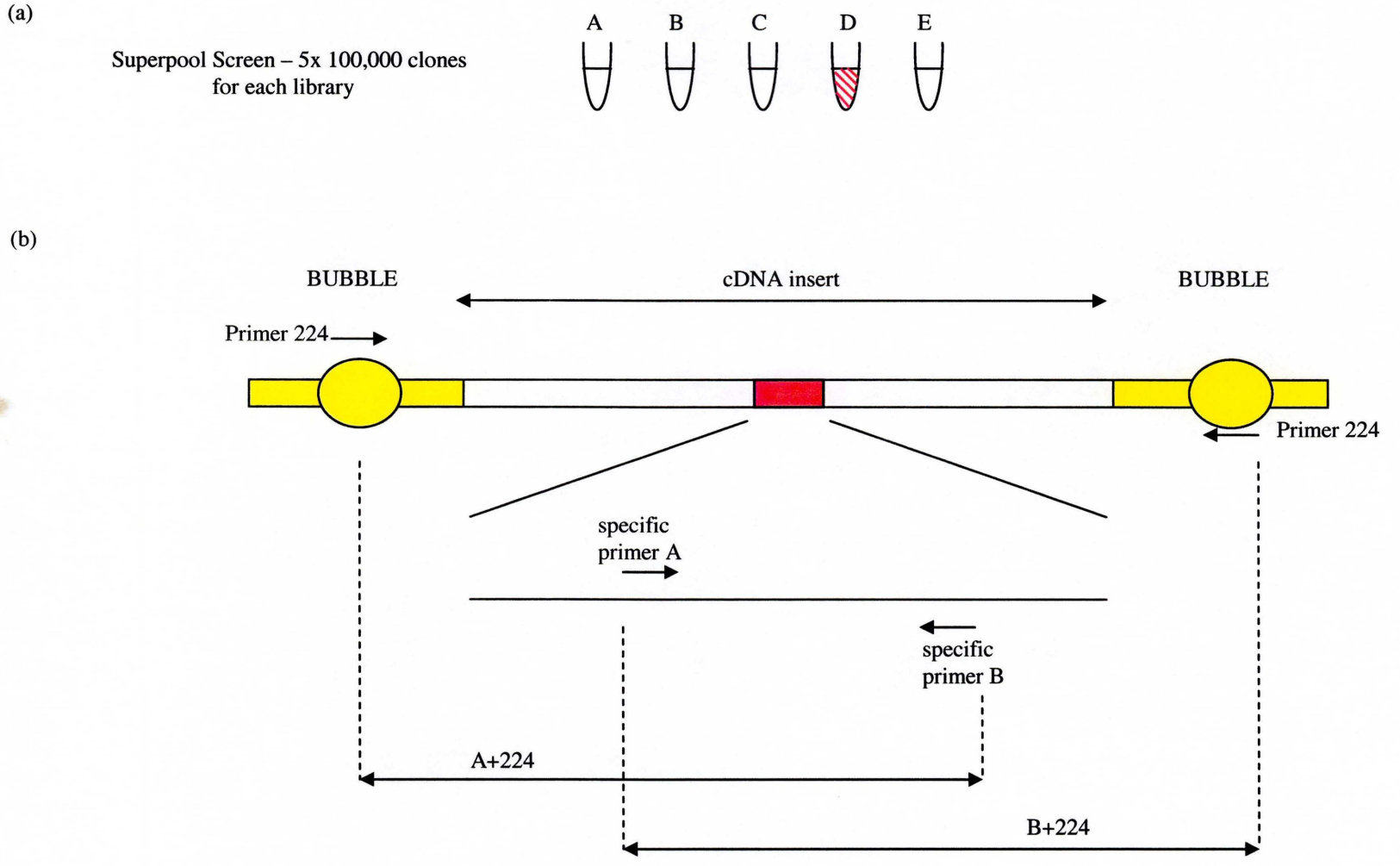
2.22.4 Vectorette PCR on cDNA libraries (illustrated in Figure 2.2)

1. Vectorette PCR was performed on the superpools of the cDNA libraries. PCR was performed using 5 µl of the diluted superpools as template in a 15 µl final volume using buffer conditions as described in Section 2.15.3. Primer combinations were as follows: 224 and specific primer A, 224 and specific primer B.
2. PCR was performed in a DNA thermocyle (Omingene) using hot start. Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95°C, followed by 17 cycles of: 94°C for 5 seconds, 65 °C for 30 seconds and 72°C for 3 minutes, followed by 18 cycles of: 94°C for 5 seconds, 65 °C for 30 seconds and 72°C for 3 minutes, followed by 72°C for 5 minutes. The PCR was paused after 4 minutes of the initial denaturation and 2 µl of Taq premix (containing 0.12 µl Amplitaq, 0.12 µl TaqExtender, 0.12 µl Perfect Match, 0.5 µl 40% sucrose + cresol red, 1.14 µl T0.1E) were added to each reaction (pipetting underneath the oil).
3. Products were separated by electrophoresis through 2.5% agarose gels and visualised by ethidium bromide staining. Products were gel purified using gel extraction kits from either GeneClean™ or Qiagen™ prior to sequencing directly.

2.22.5 Reamplification of vectorette PCR products

1. In cases where multiple bands or weaker bands were observed, bands were excised and placed in 100 µl of $T_{0.1}E$. Reamplification of each band was carried out by PCR using 5 µl $T_{0.1}E$ taken from the 100 µl containing the excised band and by adding 1.5 µl 10x NEB Buffer, 1.5 µl 5 mM dNTPs, 0.495 µl 5 mg/ml BSA, 0.21 µl 1:20 βME, 3.17 µl 40 % sucrose + cresol red, 0.375 µl 224pure primer, 0.75 µl 100ng/µl specific primer, 0.12 µl Amplitaq, 0.12 µl TaqExtender, 0.12 µl Perfect Match, 0.5 µl 40% sucrose + cresol red, 1.14 µl $T_{0.1}E$ to a well of a 96-well microtitre plate.
2. PCR was performed on a DNA thermocycler (MJ). Cycling conditions were are follows: an initial denaturation step of 5 minutes at 95°C followed by 35 cycles of 94°C for 5 seconds, 60 °C for 30 seconds and 72°C for 3 minutes, followed by 72°C for 5 minutes.
3. Products were separated by electrophoresis through 2.5% agarose gels and visualised by ethidium bromide staining. Products were gel purified using gel extraction kits (either GeneClean™ or Qiagen™) prior to sequencing directly.

Figure 2.2: *(see over) Strategy for vectorette PCR on cDNA libraries (a) Superpools representing 100,000 clones (A-E) were screened by PCR and positive super pools recorded (e.g. D, shown in red). (b)Rescue of the insert of the cDNA of interest by vectorette PCR. Insert of cDNA is shown as white rectangle, ligated 'bubble' is shown in yellow. Original position of primers for pool screening is indicated by a red rectangle. A combination of the 'bubble' primer (224) and insert specific primers (e.g A and B) were to generate two products (e.g. 224+A, 224+B) representing the entire insert of the cDNA of interest.*


(a)
Superpool Screen – 5x 100,000 clones
for each library
A
B
C
D
E
(b)
BUBBLE
cDNA insert
BUBBLE
Primer 224
Primer 224
specific
primer A
specific
primer B
A+224
B+224

2.23 Mapping and sequence analysis software and databases

2.23.1 IMAGE

All processing of fingerprinting gels was carried out using IMAGE. IMAGE processed gels from radioactive, fluorescent and *Hin*d III fingerprinting and extracted a normalised band pattern for each lane on a gel. Several procedures were run on each gel in turn:
Lane tracking – a grid was superimposed on the gel image and the grid manually edited to ensure it exactly matched the lanes on the gel.
Band calling – an analysis module traced the band pattern along the lanes and tried to identify the bands. Manual editing ensured the correct bands are chosen.
Marker locking – in order to compare band patterns from one gel to another all band positions were normalised to one master gel. A set of DNA fragments of known length or migration distance was loaded as a marker lane (see Section 2.13 for specific marker patterns for each method of fingerprinting used). Manual editing ensured the standard pattern matched to the pattern from the master gel.
Normalisation – once the marker lane patterns were locked onto the standard lane, the band positions of the sample lanes were normalised so that each lane appeared to have been run on the master gel with all distortions cancelled out. IMAGE finally generated a 'Bands' file for each gel containing normalised migration distances for all selected bands in each clone lane.

2.23.2 FPC

All contig construction described in this thesis was carried out in FPC. FPC took as the input a set of clones and their restriction fragments (called Bands) from IMAGE. Each fingerprint pattern for each clone is compared to the fingerprint patterns of all other clones in the database. The relationship between two clones was reported as a probability of coincidence, i.e. the probability that two clones overlap by chance. Two variables can be set to filter the reported overlaps:

Cut off –a match between two clones will only be reported if the probability of coincdence is less than or equal to the cut off. When analysing matches between cosmids, the tolerance was set to 1e –04, and when analysing larger insert PAC and BAC clones, the tolerance was decreased to 1e –08.)

Tolerance – two bands are considered as their migration distances differ by less than tolerance. For the analysis carried out in this thesis the tolerance was set to 7.

Overlapping clones were identified automatically and contigs were constructed manually using the available editing tools provided by FPC. Initially, two clones were positioned overlapping by the number of bands they had in common. Subsequent clones were positioned in the contig based on the number of bands they shared with the existing clones in the contig. Marker data was imported from Xace and integrated into the FPC contigs. A minimum set of clones for sequencing was chosen based on a combination of shared bands and shared marker data.

Contig Sizing –one unit in the contig display represents one fingerprint band, allowing for estimates to be obtained of contig sizes. For each method of fingerprinting, a kilobase/band figure was derived. *For radioactive fingerprinting of cosmids*, one band was the equivalent of 2 kb, based on the fact that the average size of a cosmid is 40 kb and the average number of fingerprint bands for each cosmid was 20. *For fluoresecent fingerprinting of PACs and BACs*, and average figure for each clone type was calculated based on the number of bands observed in clones whose insert sizes were known by genomic sequencing. For regions in a contig covered by PACs, a figure of 3.6 kb/band was used, and for regions of a contig covered by BACs, a figure of 4.4 kb/band was used. *For Hind* III *fingerprinting of PACs and BACs*, a figure 4.4 kb/band was used and was based on the average number of bands observed in clones whose insert sizes were known from genomic sequencing.

2.23.3 Xace

All mapping and sequencing data generated in this thesis were stored in Xace, a chromosome-specific implementation of ACeDB. ACeDB was originally developed for the *C. elegans* genome project (Richard Durbin and Jean Thierry-Mieg, 1991) A. *C. elegans* Database.

ACeDB works using a system of windows and presents data in different types of windows according to the type of data. All windows are linked in a hypertext fashion, so that clicking on an object will display further information about that object. For example, clicking on a region of a chromosome map will highlight landmarks mapping to that part of the chromosome; clicking on a landmark will display information about that landmark including landmark-clone associations, etc.

All PAC, BAC and cosmid library filters and polygrids are represented graphically in Xace. and data were entered directly. Data were then saved in the database establishing landmark-to-clone associations which can be displayed as text windows relating to either the landmark or the clone. Data can also be entered via text windows or via an internal web page. PCR library pool screening and colony PCR results were entered via the text windows.

In addition to the data generated by the X chromosome mapping group, Xace also contains displays published X chromosome maps, which have been used as part of the project. This greatly facilitates integration of maps from different sources. Genomic sequence data is also displayed in ACeDB along with the collated results from the computational sequence analysis performed by the Sanger Institute Human Sequence Analysis Group.

Xace can be accessed by following the instructions at http://www.sanger.ac.uk/HGP/ChrX.

2.23.4 Blixem

Individual matches identified as a result of similarity searches using the BLAST algorithm, or matches between sequences of cDNA clones or PCR products amplified from genomic DNA generated as part of the project, were viewed in more detail using Blixem. Blixem, (Blast matches In an X-windows Embedded Multiple alignment) is an interactive browser of pairwise Blast matches displayed as a multiple alignment. Either protein or DNA matches can be viewed in this way at either the amino acid or nucleotide level. Blixem contains two main displays: the bottom display panel shows the actual alignment of the matches to the genomic DNA sequence, and the top display shows the relative position of the sequence being viewed within the context of the larger region of genomic DNA. A program "efetch" retrieves the record from an external database (*e.g.* EMBL, SWISSPROT).

2.23.5 RepeatMasker

Human repeat sequences were masked using RepeatMasker, a program that screens DNA sequence for interspersed repeats and low complexity DNA sequence (Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). The output of the program is a detailed annotation of the repeats that are present in the query sequence. Sequence comparisons are performed by the program cross_match, an implementation of the Smith-Waterman-Gotoh algorithm developed by P. Green. The interspersed repeat datbases screened by RepeatMasker are based on the repeat databases (Repbase Update) copyrighted by the Genetic Information Research Institute.

Chapter 3

Construction of a Sequence-Ready Bacterial Clone Contig

3.1 Introduction

The generation of clone maps covering large regions of the human genome has evolved significantly over the last five years. The international collaboration to map and sequence the human genome has brought about the adaptation of existing methods and the development of new ways to generate bacterial clone maps to sequence large genomes accurately and efficiently. Some of these developments, such as bacterial clone fingerprinting were pioneered in the mapping and sequencing of small genomes (*C. elegans* Sequencing Consortium, The, 1998; Coulson, A., *et al.*, 1986; Goffeau, A., *et al.*, 1996; Olson, M. V., *et al.*, 1986). This chapter will describe the application and evaluation of large-scale mapping techniques and describe how they evolved along with the available resources during the construction of a sequence-ready bacterial clone map covering approximately 6 Mb of human chromosome Xq22 between DXS366 and DXS1230.

A 6.5 Mb YAC map was previously constructed in Xq22 between DXS366 and DXS87 (Vetrie, D., *et al.*, 1994) and included four genes and fifteen previously mapped genetic markers (Dib, C., *et al.*, 1996) (see Figure 3.1). The genes had previously been identified because of their role in a variety of diseases and included the PLP gene, defects in which cause Pelizaeus-Merchbacher Disease (PMD) (Hudson, L. D., *et al.*, 1989; Trofatter, J. A., *et al.*, 1989). There are still a number of diseases for which no gene has been cloned, but for which the critical regions include the region of interest in this chapter between DXS366 and DXS1230. For instance, genetic analysis of a family with X-linked megalocornea showed close linkage to DXS94 and DXS87 (Mackey, D. A., *et al.*, 1991).

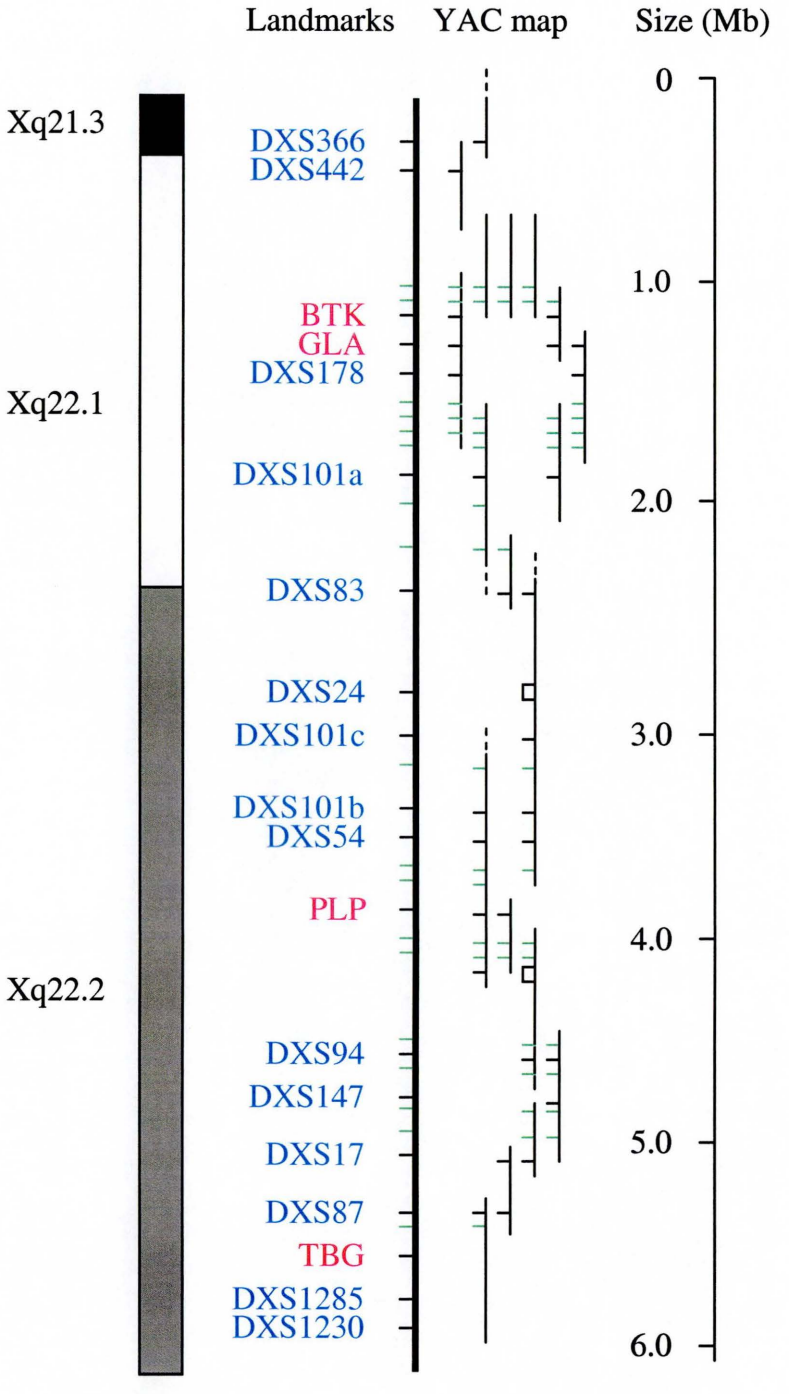


Figure 3.1: *The status of the region of interest before the generation of the bacterial clone contig began (modified from Vetrie, D., et al., 1994). The YAC map was constructed using known genes (shown in red), genetic markers (shown in blue) and additional end probes (shown as green lines). The region is approximately 6 Mb based on the sizing of YACs by pulsed-field gel electrophoresis.*

The generation of a sequence-ready bacterial clone contig in Xq22 would provide the basis for a more detailed study of the genes and sequence contained within the region, particularly with respect to the diseases for which the gene remains uncloned.

RESULTS

3.2 Contig construction

The strategy to construct a bacterial clone contig within Xq22 was based on using the available YAC map to generate initial coverage in bacterial clones (see Figure 3.2). Prior to the start of the project, a subset of the YACs from the available YAC contig (Vetrie, D., *et al.*, 1994) were pooled and used as probes by Elaine Kendall and Dave Vetrie (Guy's Hospital) to screen gridded arrays of two cosmid libraries (LLNX01 and GHc, see Section 2.7 in M&M's), which represented the best available sources of X chromosome enriched bacterial clones at the time. A subset of the YAC clones were used individually as probes to screen the cosmids. All positive cosmids were rescreened with individual YAC clones.

At the start of the project I was provided with a total of 1400 cosmids which were fingerprinted using a radioactive label, the method developed during the mapping of the *C. elegans* genome (Coulson, A., *et al.*, 1986) (see Section 2.12.1). The fingerprints for all the cosmids were digitised using IMAGE (see Section 2.23.1) and contigs were assembled using FPC (see Section 2.23.2). A total of 26 contigs covering 3 Mb or 50% of the region were generated, an example of which is shown in Figure 3.3. A summary of the status of the mapping after the cosmid fingerprinting is shown in Figure 3.16a.

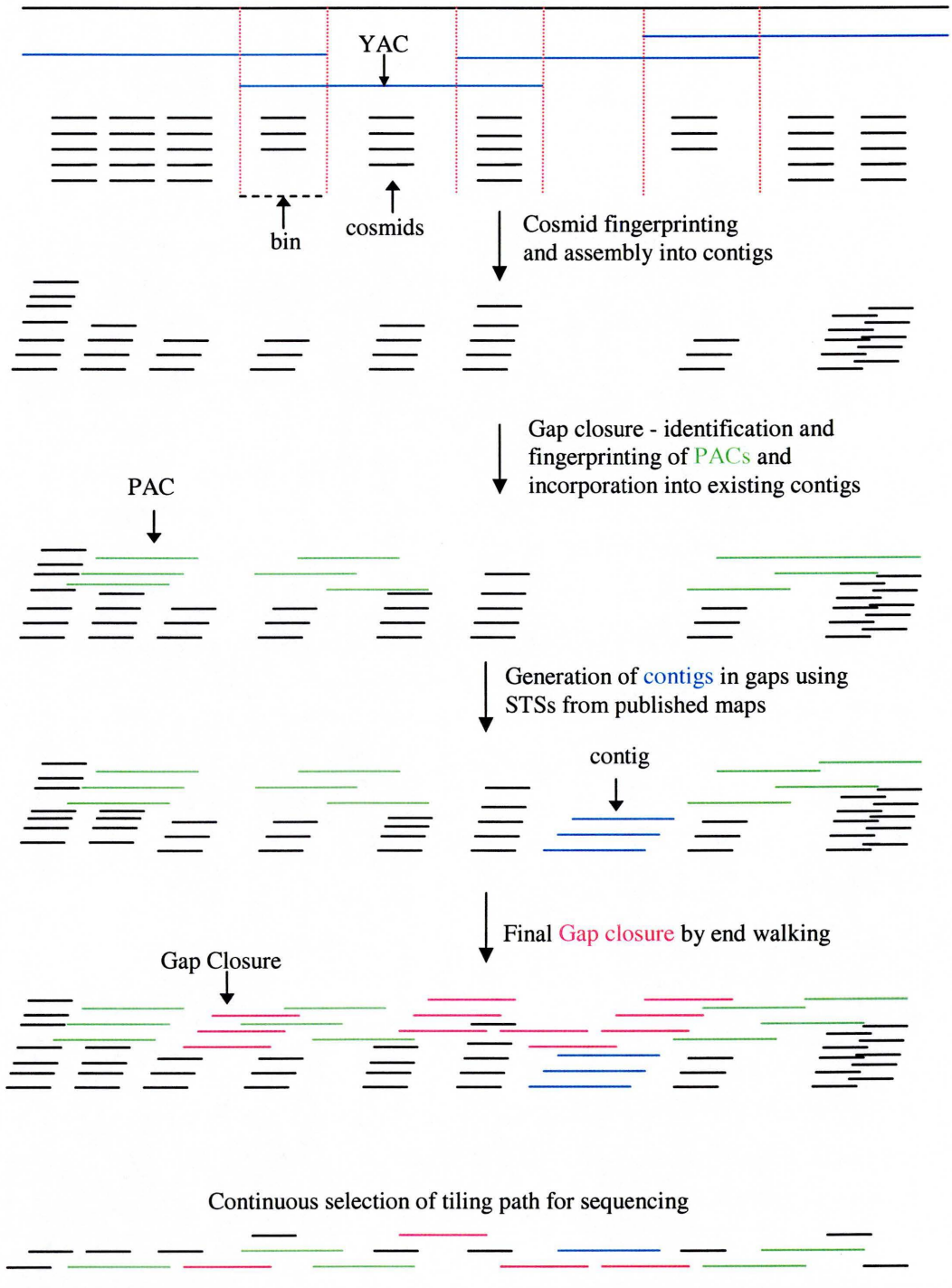


Figure 3.2: *Strategy for the construction of the bacterial clone contig. Binned cosmids are fingerprinted and assembled into contigs. Whole cosmid hybridisation identifies PACs to close gaps and extend contigs. New contigs are generated in gaps using STSs from published maps before final gap closure. At each stage, clones are chosen for genomic sequencing.*

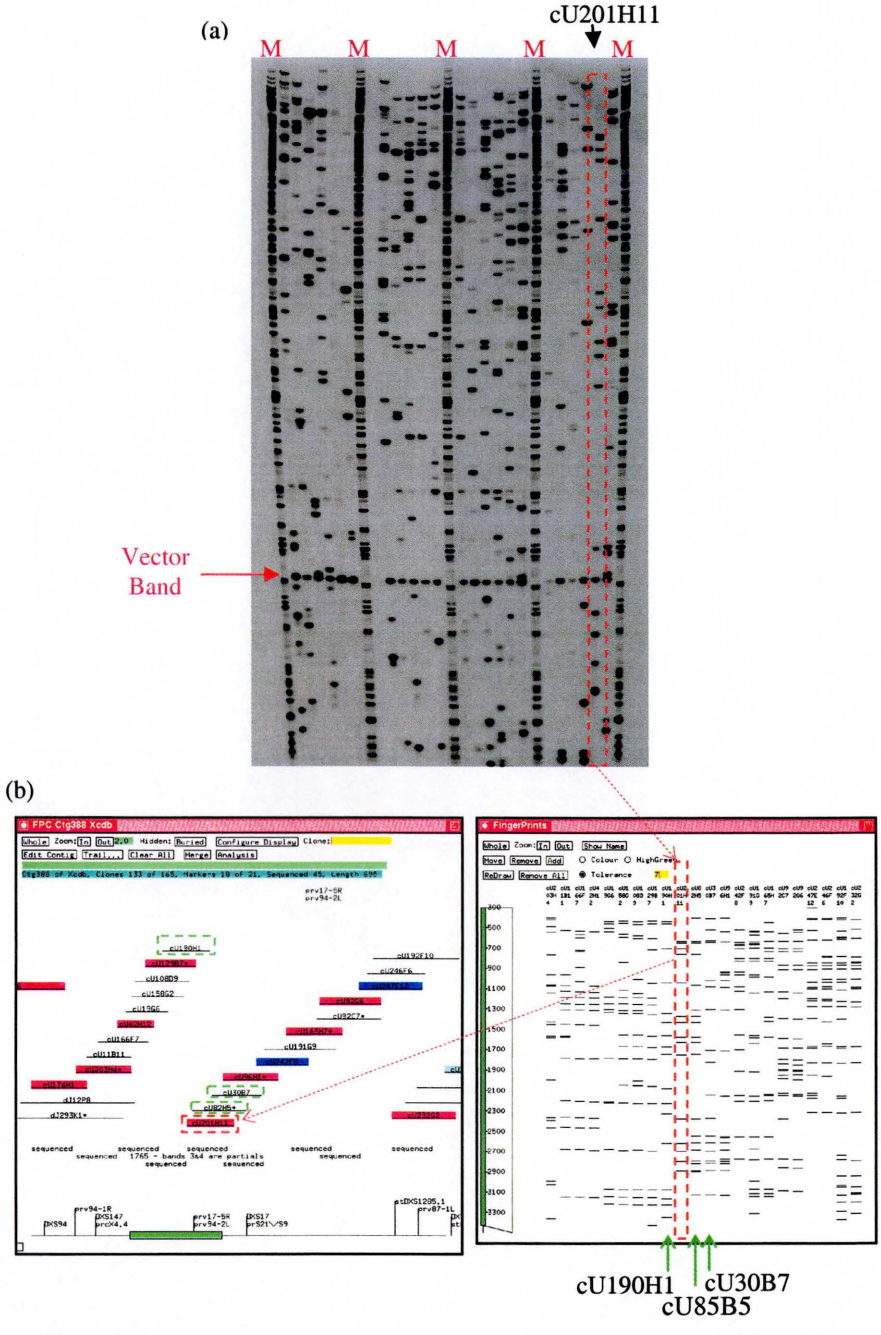


Figure 3.3: *Cosmid fingerprinting and assembly (a) An autoradiograph of the fingerprints of 24 cosmids. A marker (M) is run every seventh lane and a common vector band is seen in all sample lanes as indicated. (b) A section of one of the contigs constructed and their fingerprints. For example, lane 23 contains the fingerprint for cU201H11, which was found to overlap significantly (greater than or equal to 1e –04 - see Section 2.23.2) with cU190H1, cU82H5 and cU30B7 when compared to all other fingerprints in the FPC database.*

It had been reported that DXS101 had five loci in the region (Vetrie, D., *et al.*, 1994), two copies each at DXS101a and DXS101b, and a single copy at DXS101c. The five DXS101 loci can be distinguished by digestion of the DNA with *Eco*R1, Southern blotting of the digested DNA and probing with the DXS101 plasmid cX52.5. Each locus generates specific size fragments (5.5 kb and 7.0 kb for DXS101a, 6.0 kb and 11.5 kb for DXS101b and 13.0 kb for DXS101C). Hybridising cX52.5 (the DXS101 probe) to the available cosmids identified all those that contain the DXS101 loci (work carried out by Elaine Kendall). From the work carried out in this thesis, the fingerprinting and analysis of the DXS101-positive cosmids assembled the cosmids into three contigs. Based on the original binning of the cosmids, two of the contigs represented three of the DXS101 loci, and one contig appeared to contain both copies of DXS101 present at DXS101c (see Figure 3.4).

At this time the first of the large-insert bacterial clone libraries (PACs) became available. In order to close gaps between existing contigs, radioactively labelled *Hin*d III-digested cosmids were pooled and used as probes to hybridise to gridded arrays of PAC clones from RPCI-1 (Ioannou, P. A., *et al.*, 1994) (see Figure 3.5). A total of 149 PACs were identified with 33 cosmids. Seven cosmids failed to identify any PACs, based on the lack of overlapping PAC clones when the fingerprints were compared to those of the cosmids. It was estimated that the RPCI-1 library represented three genome equivalents and the screening carried out at this stage showed that, on average four PACs were identified with each cosmid probe, which was roughly equivalent to what was expected (three PACs for each cosmid probe).

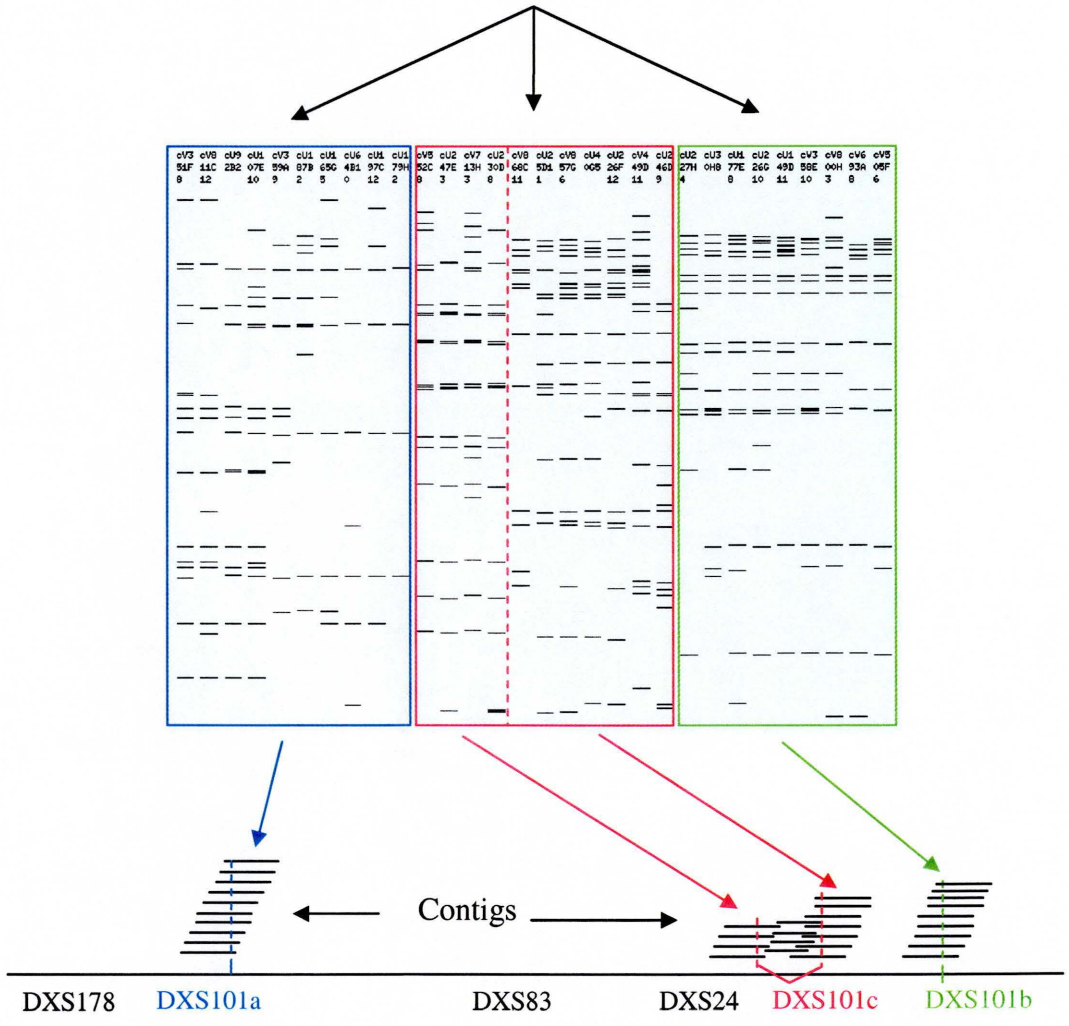
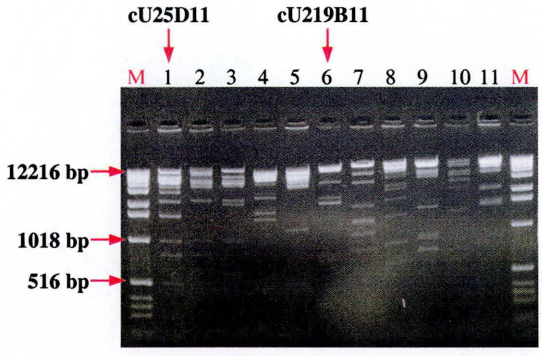


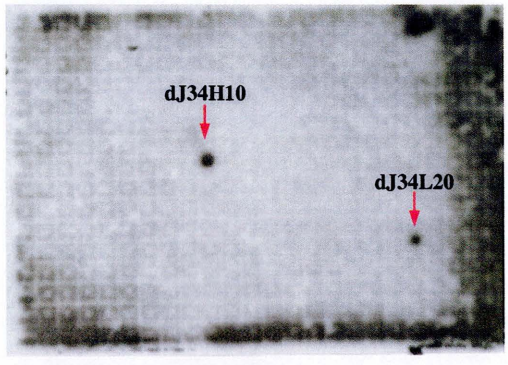
Figure 3.4: *Fingerprinting of DXS101-positive cosmids. The DXS101 positive cosmids were assembled into three contigs, based on fingerprinting, representing four of the five different loci for DXS101. Fingerprinting could not distinguish the two loci within DXS101b*

Figure 3.5*: (see over) PAC isolation by whole cosmid hybridisation (a) A photograph of a gel showing 11 cosmids digested with Hind III. Marker lanes are indicated (M). (b) An autoradiograph of one filter from the gridded PAC library showing the positive PACs identified when 6 of the 11 cosmids, including cU25D11 and cU219B11, were hybridised as a pooled probe. (c) The section of the contig showing cU25D11 overlapping with dJ34H10 and their fingerprints (dJ1055C14 and dJ593F11 were identified with an STS designed to the end of cU35G3 later in the project– data not shown). As described in Section 2.23.2, overlaps between clones are based on the number of bands they have in common. The bands in the fingerprint of dJ34H10 are shown in blue, and equivalent bands in the fingerprints of other clones are shown in red. Black bands in dJ34H10 do not match any other bands, and black bands in other clones do not match any bands in dJ34H10. The 3 black bands in cU25D11 were not seen in any other clones in the contig and were supposed to be false positives.*

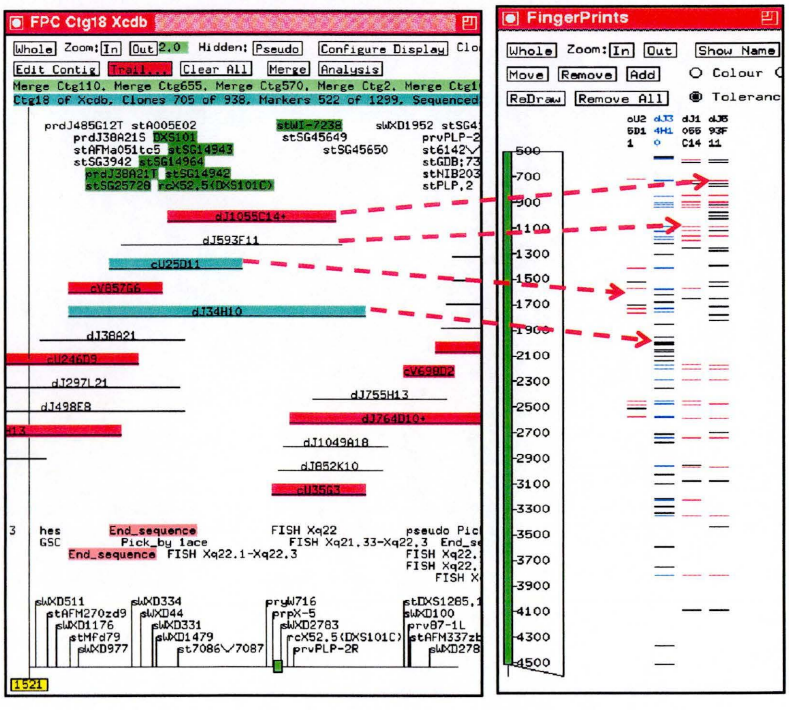
(a)


cU25D11
cU219B11
M 1 2 3 4 5 6 7 8 9 10 11 M
12216 bp
1018 bp
516 bp

(b)


dJ34H10
dJ34L20

(c)

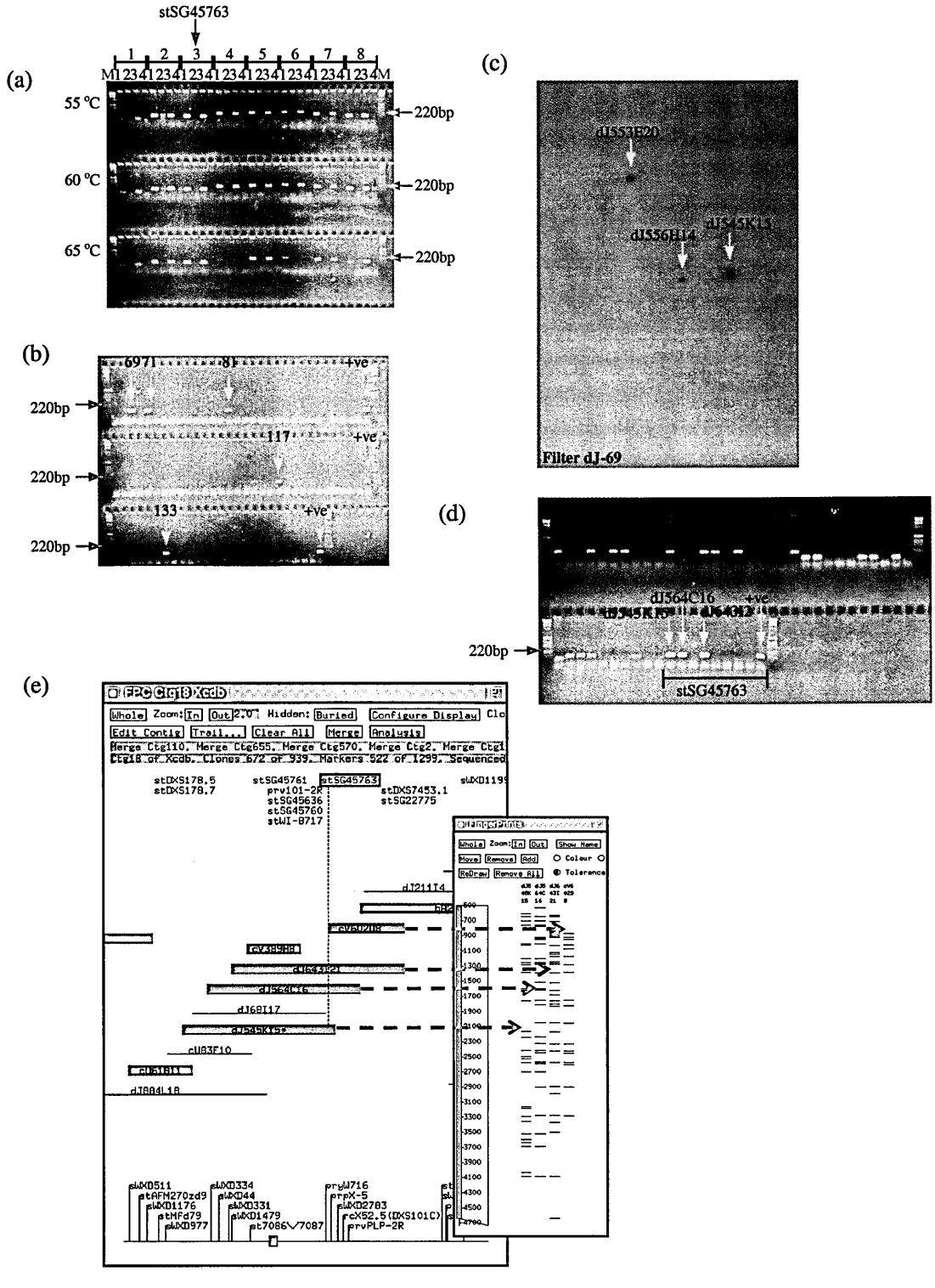

FPC Ctg18 Xcdb
FingerPrints
Whole
Zoom: In Out
Show Name
Move
Remove
Add
Colour
ReDraw
Remove All
Tolerance
Edit Contig
Trail...
Clear All
Merge
Analysis
Configure Display
Merge Ctg110, Merge Ctg655, Merge Ctg570, Merge Ctg2, Merge Ctg1
Ctg18 of Xcdb, Clones 705 of 938, Markers 522 of 1299, Sequenced
dJ1055C14
dJ593F11
cU25D11
cV857G6
dJ34H10
dJ36A21
cU24609
dJ297L21
dJ498E8
dJ755H13
dJ764D10
dJ1049A18
dJ852K10
cU3563
End_sequence
FISH Xq22
FISH Xq22.1-Xq22.3
FISH Xq21.33-Xq22.3
stWXD511
stWXD334
stWXD1952
stSG45649
stSG45650
stSG25720
stSG14943
stSG14964
stSG14942
stWXD2783
stWXD1479
stWXD977
stWXD1176
stWXD44
stWXD331
st7086
st7087
stWXD100
stWXD278
1521

The PACs were fingerprinted and 80 of the 149 (55%) were incorporated into the existing contigs thus closing fifteen gaps and extending seven ends of contigs. At the end of this second stage there were ten contigs covering 3.9 Mb or 65% of the region (see column (b) of Figure 3.16). At this time, fingerprinting by fluorescent labelling was developed (Gregory, S. G., *et al.*, 1997). In order to benefit from the increased throughput and safety of this technique, a subset of clones (all cosmids identified for sequencing and all PAC clones) were fingerprinted using this method and assembled into the same set of contiguous blocks. Other cosmids that were radioactively fingerprinted and assembled into contigs, but did not form part of the minimum set chosen for genomic sequencing, were not re-fingerprinted.

At this point in the project, the clones which extended the ends of the contigs furthest were the larger insert PAC clones, and whole cosmid hybridisation was therefore no longer useful to identify further bacterial clones for gap closure. Hybridisation of probes derived from whole PAC clones (similar to whole cosmid hybridisation) to the PAC library would have been problematical given the cross hybridisation of vector DNA between probe and target clones. Insert-specific amplification, e.g *Alu* PCR, is limited by the fact that only a fraction of the insert is amplified. At this time an STS-based YAC map was published across the region (Srivastava, A.K., *et al.*, 1999), containing additional STSs that had not been available previously. Twenty-seven STSs thought to lie in gaps between existing bacterial clone contigs (see Table 2.4) were screened against the sections of the PAC library (RPCI-1, 2, 3). This identified 50 new PAC clones which were fingerprinted and assembled into three new contigs. An example of clone isolation using these novel STSs is shown in

Figure 3.6 (see also Figure 3.16c). At the end of the third stage there were 16 contigs covering 5 Mb or 80 % of the region.

Figure 3.6: *(see over) PAC isolation using STSs taken from YAC map of Srivastava, A. K., et al. (1999) (a) Eight STSs (1-8) designed from sequence generated at the ends of 10 clones were tested for their ability to amplify unique sequence in human genomic DNA at three different temperatures of the PCR. Templates included human DNA (1), X-chromosome hybrid (2), hamster genomic DNA (3) and $T_{0.1}E$ (4). (b) One of the STSs, stSG45763 designed to one end of cV602D8 (see Table 2.5) was used to amplify DNA of pools 67-150 (each containing 2912 PAC clones) from RPCI-3 library. Five positive pools were detected, as indicated. A positive control (human genomic DNA) was run in parallel. (c) The product of amplification of genomic DNA using stSG45763 was labelled, pooled with 9 other products, and used as a hybridisation probe to screen gridded filters representing clones of each pool (1 filter represents 1 pool). The filter shown represents pool 69, and 3 positives were identified as marked. (d) Positive clones detected on filters representing the pools shown positive in (b) were streaked and individual colonies tested against stSG45763. The two other positives on filter 69 (dJ556H14 and dJ553E20) were found to be positive for another unrelated STSs. The other positive clones were identified from other filter hybridisations. (e) The fingerprints of the 3 clones show good correspondence of fingerprint patterns confirming overlap and were integrated into the contig by comparison with other fingerprints previously in the database (e.g cV602D8).*


stSG45763
(a)
1 2 3 4 5 6 7 8
M1 23 41 23 41 23 41 23 41 23 41 23 41 23 41 23 4M
55 °C
60 °C
65 °C
220bp
220bp
220bp
(b)
220bp
220bp
220bp
+ve
+ve
+ve
(c)
dJ553E20
dJ556H14
dJ545K15
Filter dJ-69
(d)
dJ564C16
+ve
220bp
stSG45763
(e)
Whole
Edit Contig
Merge
Analysis
stDXS178.5
stDXS178.7
stSG45761
stSG45763
prv101-2R
stSG45636
stSG45760
stWI-8717
stDXS7453.1
stSG22775
dJ211I4
dJ564C16
dJ68I17
dJ545K15c
cUB3E10
dJ884L18
stDXS1191
stDXS7453.1
stSG22775
stDXS178.5
stDXS178.7
stSG45761
stSG45763
prv101-2R
stSG45636
stSG45760
stWI-8717

Closure of the remaining gaps was completed using either probes generated by vectorette end rescue (see Sections 2.19 and 2.20) from the ends of clones at the ends of contigs, or STSs generated after directly sequencing the ends of the cloned PAC inserts (sequencing was carried out by Elizabeth Huckle). DNA of clones chosen for vectorette end rescue was prepared using a standard alkaline lysis and subsequent phenol chloroform extraction, and digested with *Rsa*I (see Section 2.11.1). Vectorette 'bubbles' (see Table 2.3) were ligated on to the ends of the restriction fragments and amplification of each end of the insert of each clone was carried out using vector specific primers. The amplification products were resolved on agarose gels, excised and stored to be used as templates for probe generation.

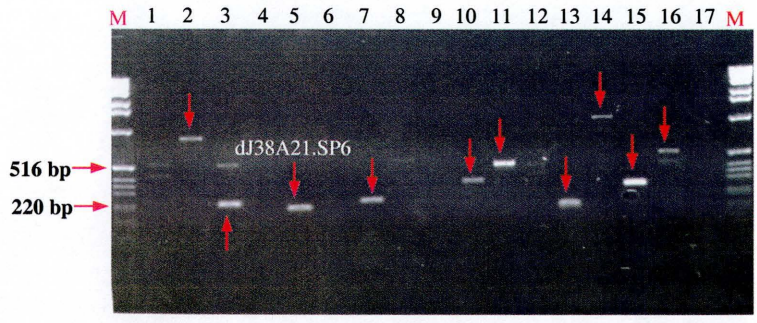
The vectorette probes or STSs from end sequencing were used to screen six bacterial clone libraries that were now available (RPCI-1, 3, 4, 5, 11 and 13) (see Figure 3.7). In total, 131 PACs and 101 BACs were identified using 20 probes and 103 STSs. All newly identified clones were fingerprinted and incorporated into the contigs and all remaining gaps were closed.

The final bacterial clone contig covers approximately 6 Mb of Xq22 between DXS366 and DXS1230 and contains 92 cosmids, 211 PACs and 101 BACs (see Figure 3.8). A total of 44 probes (24 positioned by Elaine Kendall and Dave Vetrie, 20 vectorette end probes positioned duing this project) and 130 STSs (103 STSs designed to sequence generated at the end of the clones, 27 STSs from YAC map of Srivastava, A.K., *et al.*, 1999) have been used in the construction of the contig. A minimum set of clones from the contig was chosen for genomic sequencing (carried

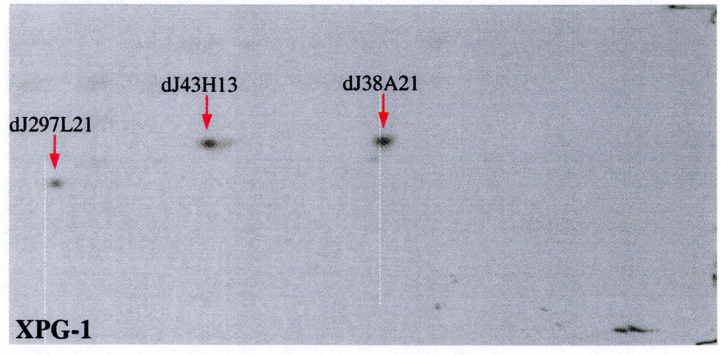
out by the Sanger Centre sequencing teams) and resulted in one contiguous segment of sequence covering 6.0 Mb.

Figure 3.7: *(see over) PAC isolation using STSs generated by vectorette PCR or end sequencing (a) Generation of 10 products from the SP6 ends of 17 PAC clones by vectorette PCR (successful amplification is indicated by an arrow). Lanes 3 and 16 contain two bands, in each case the stronger one was excised and used to generate a probe for walking. (b) One product, dJ38A21.SP6, was labelled and used as a hybridisation probe to screen two filters representing an X chromosome-specific collection of PAC and cosmid clones (XPG-1 and XPG-2). 6 positive clones were identified, 3 of which were present on the filter shown and are indicated. (c) The fingerprints of the 6 clones show good correspondence of fingerprint patterns confirming the overlap. The 5 clones (highlighted in green) were integrated into the bacterial clone map by comparing the fingerprints with other fingerprints previously in the database (e.g dJ1143D15).*

(a)


M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 M
dJ38A21.SP6
516 bp
220 bp

(b)


dJ43H13
dJ38A21
dJ297L21
XPG-1

(c)

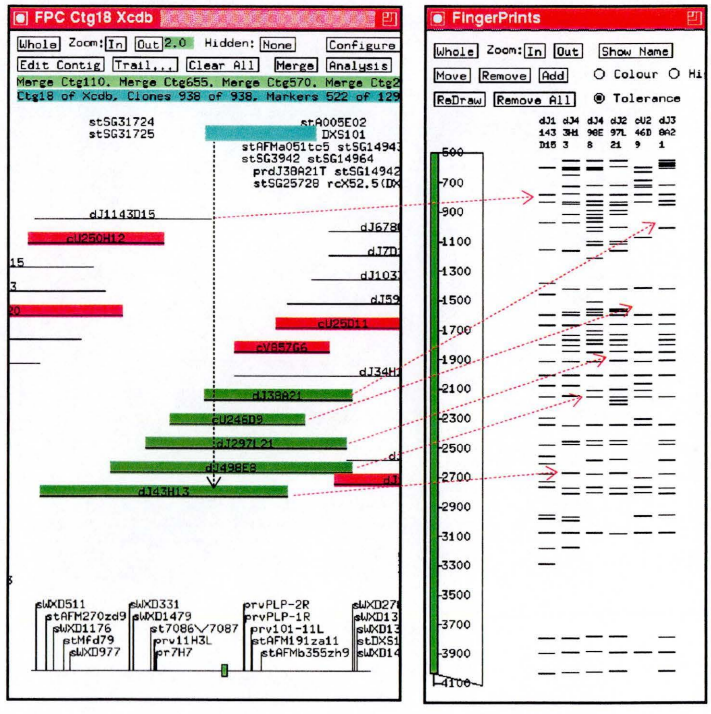
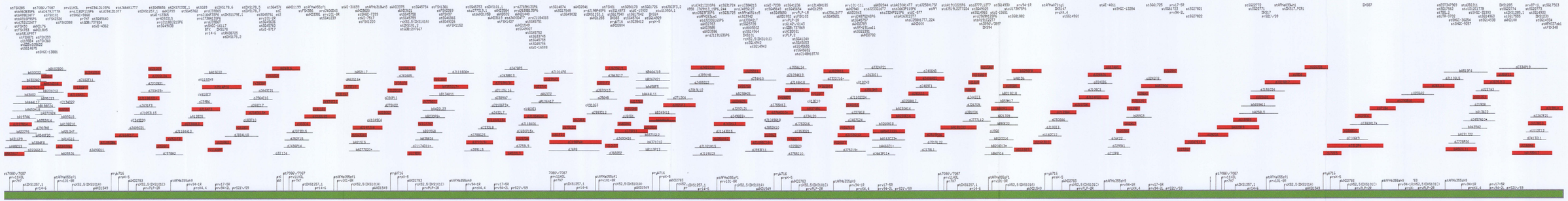

FPC Ctg18 Xcdb
FingerPrints
Whole
Zoom: In Out
Hidden: None
Configure
Edit Contig
Trail...
Clear All
Merge
Analysis
Show Name
Move
Remove
Add
Colour
ReDraw
Remove All
Tolerance
dJ1143D15
cU250H12
cU25D11
dJ43H13

Figure 3.8: *(see over) FPC diagram of bacterial clone contig between DXS366 and DXS1230. The markers used to identify the clones and confirm contig order during earlier stages of the project are shown at the top. A subset of the markers, chosen as framework markers are indicated at the bottom. Clones in the contig are indicated by horizontal black lines, the length of the clone is determined by the number of bands in the fingerprint. The overlap between clones is determined by the number of bands pairs of clones have in common. The clones shown in red were identified for the minimum tiling path for genomic sequencing.*



3.3 Comparison of the published maps

The generation of the complete sequence of the region of interest allows for the study of the accuracy of previously published maps in order to verify marker order and placement and identify conflicts with the final sequence map. It also allows for a comparison of physical distance with genetic distances. There are two types of maps available covering the region of interest between DXS366 and DXS1230: a genetic map (Dib, C., *et al.*, 1996) and three physical maps; the RH map (electronic version released in 1999, updated from Deloukas *et al* (1998) and two YAC/STS based-map (Srivastava, A. K., *et al.*, 1999, Vetrie, D., *et al.*, 1994). All STSs on these published maps have been accurately positioned on the sequence and the order and distance from neighbouring STSs compared.

3.3.1Genetic Map

The order and physical distances of markers on the final sequence map were compared with the order and genetic distances of markers on the available genetic map (Dib, C., *et al.*, 1996) (see Figure 3.9). The region between DXS366 and DXS1230 contains eight genetic markers that are placed within 2.5 cM of each other on the genetic map. Three markers (shown in grey in Figure 3.9), placed within the same region on the genetic map could not be identified in the sequence between DXS366 and DXS1230. All available X chromosome finished and unfinished sequence was searched using BLAST (Altschul, S. F., *et al.*, 1990) and the results revealed that two markers (AFMb083yb5 and AFMa052xc1) are located approximately 500 kb distal to DXS1230 and the third (AFMa162yc9) is located approximately 4 Mb distal to DXS1230.

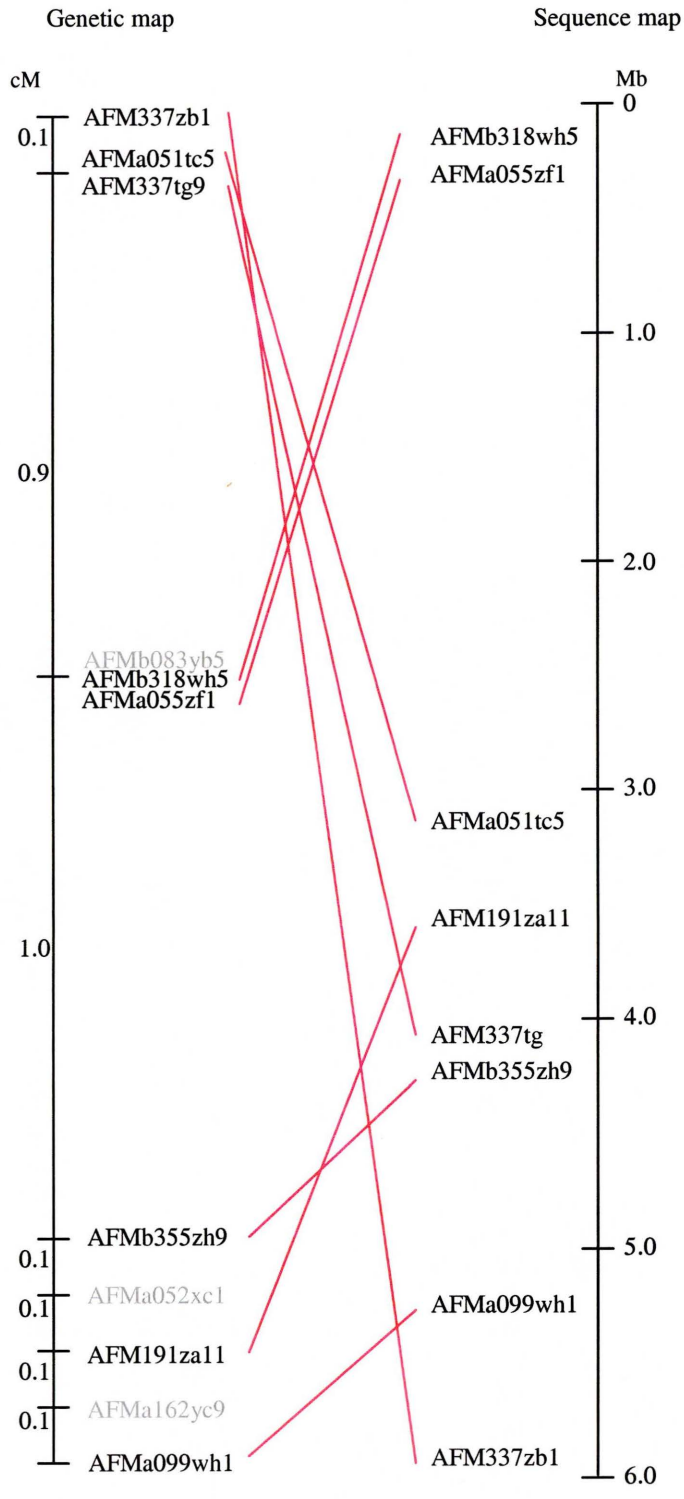


Figure 3.9: *Comparison of the genetic map. A comparison of the STS order and genetic versus physical distance, between the published genetic map (on the left) and the final sequence map (on the right). Red bars link the same genetic marker and marker names in grey have not been able to be placed within the final sequence map.*

There is some disagreement between the order of the eight genetic markers identified in the sequence and their order on the genetic map. For instance, AFMb318wh5 and AFMa055zf1 are placed 0.9 cM distal to AFM337zb1, AFMa051tc5 and AFM337tg9, but are 3 Mb more proximal in the sequence. Analysis of the draft sequence showed that for long chromosome arms 1 cM equals approximately 1 Mb whereas for the shortest chromosome arms 2 cM equals approximately 1 Mb. Recombination is also not uniform across these regions. There are regions where recombination is less frequent (e.g. towards centromeres) and other regions that appear to have a higher recombination frequency (e.g. towards telomeres) (IHGSC, 2001). Three markers (AFM337zb1, AFMa051tc5 and AFM337tg9) have been placed within 0.1 cM on the genetic map, but cover a distance of 3 Mb on the sequence map, which may represent a region of low level recombination. Although the genetic mapping has been able to cluster the eight markers in one region of the genome, it has not been able to identify their correct order. The eight markers have been positioned on the genetic map with odds of greater than 1000:1 that there is no other more likely position. However, the markers are all positioned within a 3 cM interval, which is reaching the limitations of resolution for genetic mapping. This may account for the differences in the marker order based on genetic mapping, and the actual marker order identified from the sequence.

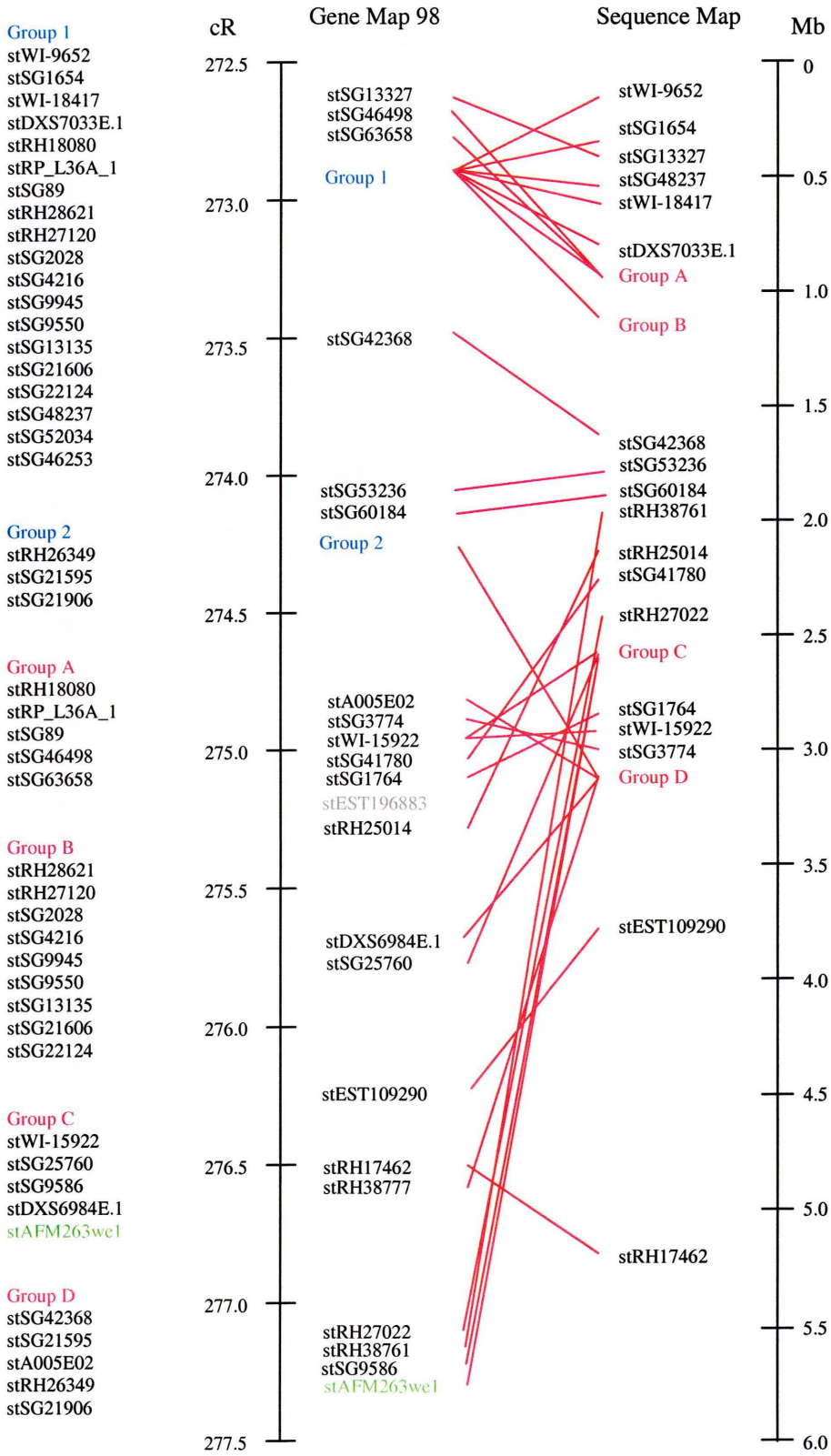
3.3.2 RH Map

The order and physical distances (given in centiRays (cR) of markers placed on the RH map were compared to the order and distance of markers on the final sequence map (see Figure 3.10). The region of interest between DXS366 and DXS1230 is located within one bin of the RH map between the two framework markers DXS990

and DXS1106 (the STS for DXS1106, stAFM263we1, is positioned within the region and is shown in green in Figure 3.10). The 6 Mb region is estimated to be 5 cR according to the RH map, which is similar to the reported figure that on average, 1 cR is equivalent to 1 Mb (Deloukas, P., *et al.*, 1998). A total of 44 markers have been positioned within the region of interest by RH mapping and 43 of them have been located within the final sequence map (see Figure 3.10). The sequence for the one marker that has not been placed in the final sequence map (stEST196883 – shown in grey on Figure 3.10) did not match any human genome sequence currently available. In one case, stWI-15922 appears once on the RH map but twice on the final sequence map.

The main difference between the two maps lies between 275.0 cR and 277.5 cR on the Gene Map (a region of 2.5 cR) and between 2.5 Mb and 3.0 Mb on the sequence map (a region of 0.5 Mb) where there is also some discrepancy between the marker order and marker distances. The markers in the region of RH map between 275.0 cR and 277.5 cR are clustered within the 0.5 Mb region of the final sequence map.

Figure 3.10: *(see over) Comparison of the gene map. A comparison of the STS order and centi-Ray versus physical distance between the published RH Map (on the left) and the final sequence map (on the right). Red lines link the same markers. Names in grey indicate those markers that could not be placed within the final sequence map. Groups 1, 2, A, B, C and D represent clusters of markers positioned too closely on either the RH map or in the sequence map to be resolved on the diagram.*


Group 1
stWI-9652
stSG1654
stWI-18417
stDXS7033E.1
stRH18080
stRP_L36A_1
stSG89
stRH28621
stRH27120
stSG2028
stSG4216
stSG9945
stSG9550
stSG13135
stSG21606
stSG22124
stSG48237
stSG52034
stSG46253
Group 2
stRH26349
stSG21595
stSG21906
Group A
stRH18080
stRP_L36A_1
stSG89
stSG46498
stSG63658
Group B
stRH28621
stRH27120
stSG2028
stSG4216
stSG9945
stSG9550
stSG13135
stSG21606
stSG22124
Group C
stWI-15922
stSG25760
stSG9586
stDXS6984E.1
stAFM263we1
Group D
stSG42368
stSG21595
stA005E02
stRH26349
stSG21906
cR
272.5
273.0
273.5
274.0
274.5
275.0
275.5
276.0
276.5
277.0
277.5
Gene Map 98
stSG13327
stSG46498
stSG63658
Group 1
stSG42368
stSG53236
stSG60184
Group 2
stA005E02
stSG3774
stWI-15922
stSG41780
stSG1764
stEST196883
stRH25014
stDXS6984E.1
stSG25760
stEST109290
stRH17462
stRH38777
stRH27022
stRH38761
stSG9586
stAFM263we1
Sequence Map
stWI-9652
stSG1654
stSG13327
stSG48237
stWI-18417
stDXS7033E.1
Group A
Group B
stSG42368
stSG53236
stSG60184
stRH38761
stRH25014
stSG41780
stRH27022
Group C
stSG1764
stWI-15922
stSG3774
Group D
stEST109290
stRH17462
Mb
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0

3.3.3 YAC maps

There are two published YAC maps that include the region of interest between DXS366 and DXS1230. As mentioned in the introduction to this chapter, one of the published YAC maps was used as the basis for generating the initial coverage in bacterial clones (Vetrie, D., *et al.*, 1994). The marker order in this YAC map was consistent with the order on the final sequence map (data not shown). Although the markers in the YAC map were used to identify and orient bacterial clone contigs during the construction of the final sequence-ready contig, the order of the markers was confirmed independently through genomic sequencing and there are no inconsistencies between the bacterial clone contig and the sequence map.

The order of the markers on the second published YAC map (Srivastava, A. K., *et al.*, 1999) was compared to the order in the final sequence (see Figure 3.11). There are four regions that show inconsistencies, primarily through inversions of groups of markers. Although the order generated by Srivastava, A.K., *et al.* differed from the final order in the sequence map, the YAC map was a valuable resource during this project as a source of STSs.

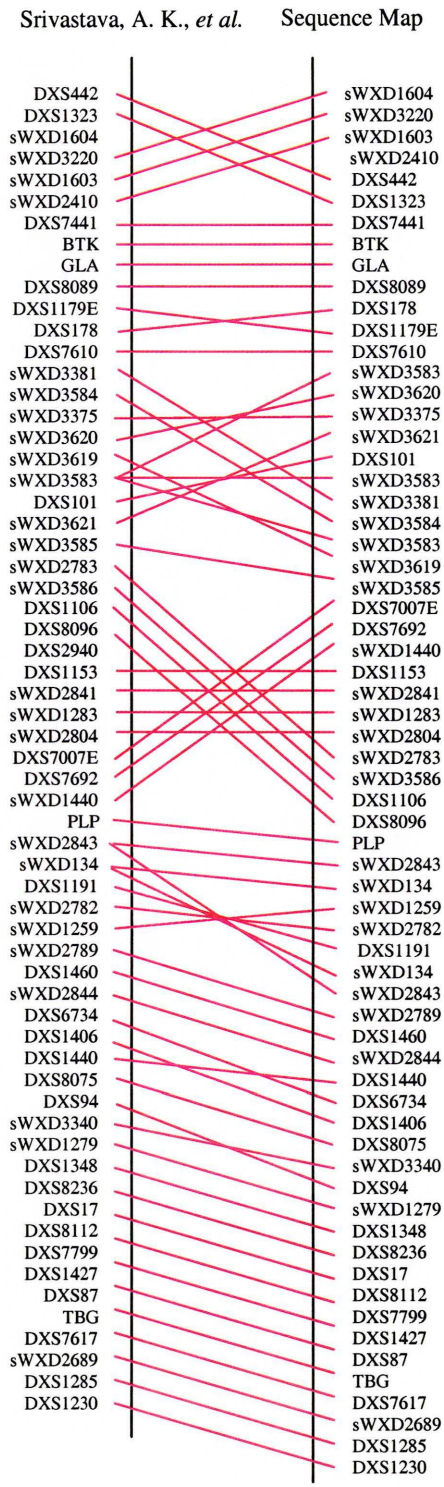


Figure 3.11: *Comparison of the YAC map. A comparison of STS order between the published STS-based YAC contig of Srivastava, A.K., et al (1999)(on the left) and the final sequence map (on the right). Red lines link the same marker.*

3.4 Sequence composition and repeat content analysis

The complete sequence provides the opportunity to analyse the base composition and the repeat of the region, and in some cases identifying specific sequences that generated conflicting data which required resolution before the completion of the sequence.

3.4.1 Sequence composition analysis

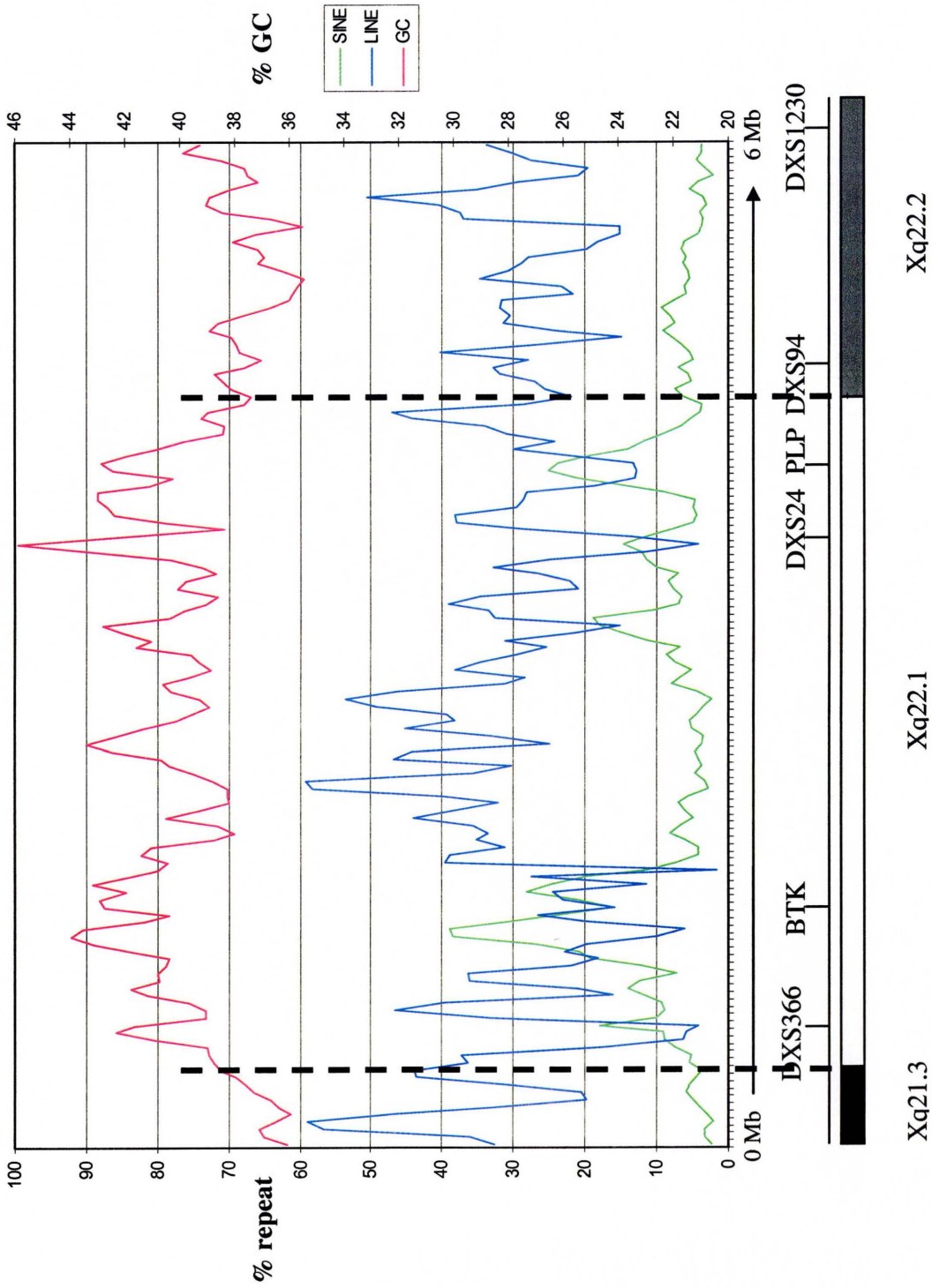
The boundaries of the contig constructed in this chapter are identified as being DXS366 at the proximal end and DXS1230 at the distal end. DXS366 was identified as a genetic marker (variable number of tandem repeats – VNTR) and originally placed broadly on the cytogenetic map between Xq21.2 and Xq24, based on a series of X chromosome translocation breakpoints (Dietz-Band, J. N., *et al.*, 1990). It was more recently placed in proximal Xq22 by YAC mapping (Vetrie, D., *et al.*, 1994). DXS1230 was also identified as a genetic marker (dinucleotide repeat) and had been localised to approximately 6 Mb distal to DXS366 by YAC mapping (Vetrie, D., *et al.*, 1994). Cytogenetic bands are sized as a fraction of the total length of the chromosome. In the case of Xq22.1, it is estimated to be approximately 5 Mb given the X chromosome is 164 Mb in size and Xq22.1 is approximately 32 times smaller than the total length of the X chromosome. Based on these size estimates, this would place DXS1230 in Xq22.2.

The evidence for the localisation of the flanking markers DXS366 and DXS1230 would suggest that the contig described in this chapter spans part of Xq22.1 and part

of Xq22.2. Xq22.1 is a light band, whereas Xq22.2 is a dark band. Dark bands are associated with high AT (or low GC) due to the fact that certain chromosome stains such as DAPI bind AT rich regions preferentially (Schnedl, W., *et al.*, 1977) and subsequently, light bands are associated with lower AT (or higher GC).

In order to analyse the sequence content in the region of interest between DXS366 and DXS1230, the available genomic sequence from approximately 650 kb upstream of DXS366 to DXS1230 was divided into 100 kb segments, overlapping by 50 kb. The GC content of each 100 kb segment was then analysed using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) and each result plotted as single point on a linear scale. In order to identify any correlation between GC content and repeat content in the region the SINE content and LINE content of the same 100 kb segments of sequence were also analysed (see Figure 3.12).

Figure 3.12: *(see over) A graph showing the relative abundance of the GC content (red), LINES (in blue) and SINES (in green) across the region of interest. The scale for GC content is given on the right side of the graph, and the scale for SINE and LINE content is given on the left side of the graph. The position of markers previously placed on the cytogenetic map is also indicated.*


% GC
SINE
LINE
GC
% repeat
0 Mb
6 Mb
DXS366
BTK
DXS24
PLP
DXS94
DXS1230
Xq21.3
Xq22.1
Xq22.2

The GC plot shows that for the region predicted to be within Xq22.1, the GC content remains above 38%, whereas the regions flanking Xq22.1, namely Xq21.3 and Xq22.2 (both dark bands) the GC content drops below 38%. This is consistent with the assumption that light bands are GC richer than dark bands. Xq22.1 is consistently higher than the genome average of 41% (figure taken from the analysis of the draft sequence (IHGSC, 2001). In general, Xq21.3 and Xq22.2 appear to be higher in LINE and lower in SINE, but the SINE and LINE content of Xq22.1 is much more variable.

3.4.2 Analysis of previously identified low copy repeats

It is well known that common repeats such as SINES are widely dispersed in the human genome. As discussed in Section 3.2, it had been reported that there were five copies of DXS101, a low copy repeat specific to Xq22. The five copies had been placed in three different loci, DXS101a, DXS101b and DXS101c (DXS101a and DXS101b each contain two copies of DXS101) (Vetrie, D., *et al.*, 1994). The sequence of each locus has not previously been determined and the probe used to identify the DXS101-positive cosmids was not available for this project. In order to identify the positions of DXS101 within the sequence, genomic sequences thought to include each copy of DXS101 (based on previous hybridisation to the available cosmids carried out by Elaine Kendall) were compared to each other by BLAST. Five regions of approximately 900 bp have been identified that appear to represent the previously reported DXS101 repeat. The results are summarised in Table 3.1.

Table 3.1: *Position of the DXS101 loci in the genomic sequence*

DXS101 Locus	Genomic sequence	Position in sequence	Size (bp)	Reported Restriction Fragment Size (kb)	Actual Restriction Fragment Size (kb)
DXS101a_1	dJ122O23	21992-22909	917	7.0	7.5
DXS101a_2	cV351F8	12407-13313	907	5.5	5.5
DXS101b_1	cV857G6	16018-16898	881	11.5	11.6
DXS101b_2	cV857G6	38408-39297	890	6.0	6.3
DXS101c	cU177E8	35575-36468	894	13.0	12.9

Each of the five sequences identified were located within *Eco*RI restriction fragments (see actual restriction fragment size in Table 3.1) that corresponds to those identified previously when the DXS101 probe was hybridised to *Eco*RI digested DXS101-positive YAC clones (Vetrie, D., *et al.*, 1994) (see reported restriction fragment size in Table 3.1). The five sequences were aligned using CLUSTALW and appear to cluster into two groups (data not shown). Group 1 contains DXS101a_1, DXS101b_1 and DXS101b_2 and are greater than 85 % identical to each other. The second group contains DXS101a_2 and DXS101c and are 80 % identical to each other. Within the 900 bp there is a region of approximately 100 bp that is greater than 90 % identical in all five sequences and would account for the ability to identify all loci by hybridisation with the DXS101 probe.

3.4.3 Analysis of previously unidentified low copy repeats

The sequence of the region allows for previously unidentified low copy repeats to be characterised. The 6 Mb of sequence was analysed for repeats using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) to remove previously

characterised common repeats and compared to itself by BLAST. All self-matches were removed and the remaining results viewed in ACT (see http://www.sanger.ac.uk/Software/ACT) (see Figure 3.13). The results are summarised in Table 3.2. There are six duplications greater than one kb in length which are greater than or equal to 99% percentage identical.

Table 3.2: *Low copy duplications between DXS366 and DXS1230*

Repeat	Type	Length (kb)	Identity	Starting Positions in sequence map (kb)
1	Inverted	5	99	1168 1187
2	Inverted	1	100	1411 1891
3	Direct	1	99	1412 1939
4	Inverted	140	99	1769 1920
5	Direct	18	99	3518 3536
6	Inverted	12	100	5804 5838

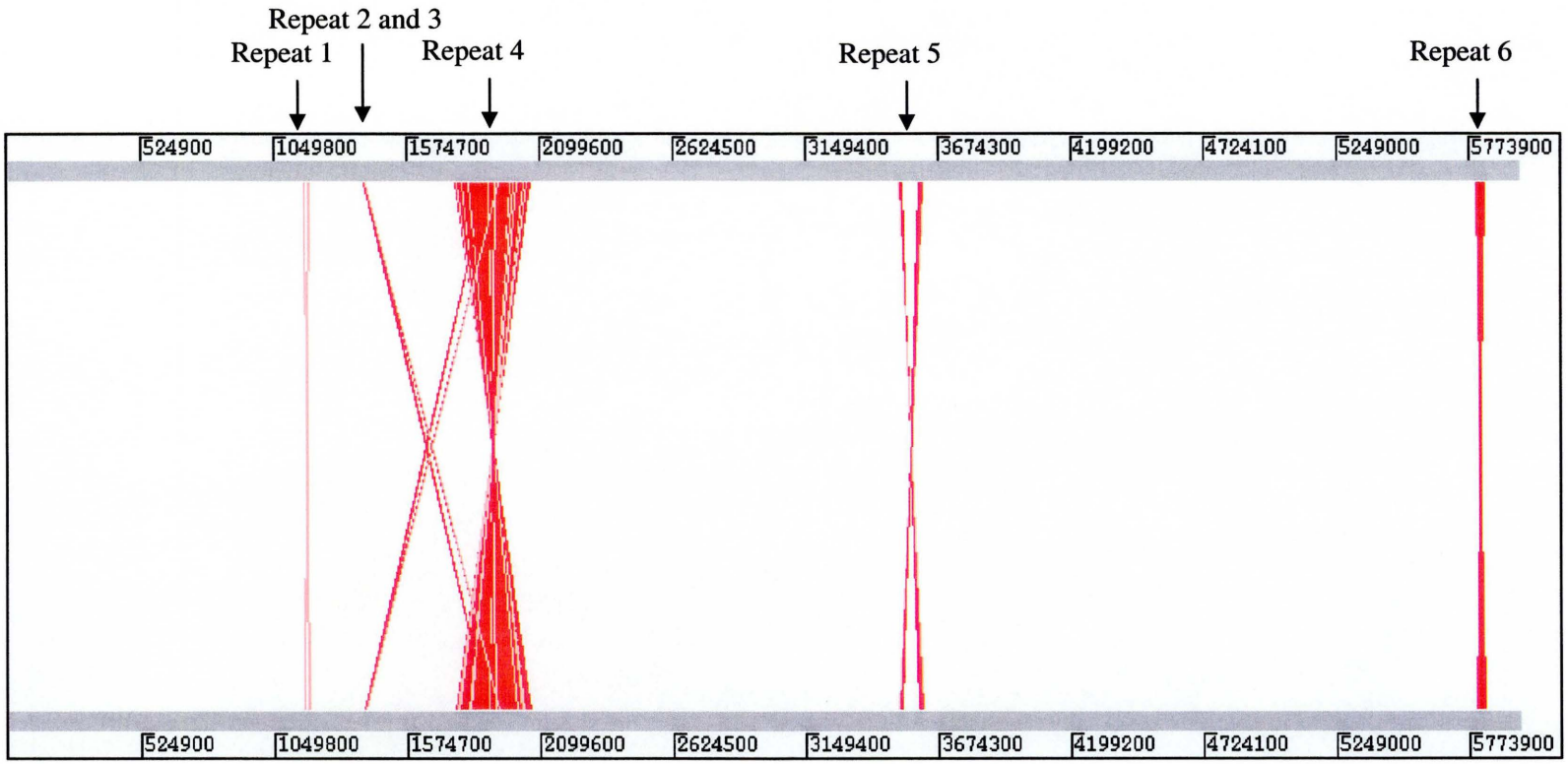
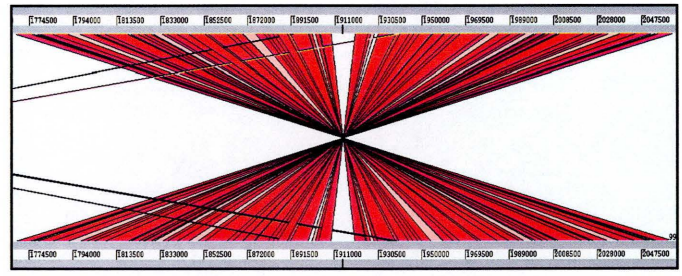


Figure 3.13 *An image from the computer program ACT, showing the position of low copy repeat sequences within the final sequence map. The grey numbers indicate a base pair scale. The region was compared to itself by BLAST, and red bar links similar sequences. The region is represented twice, along the top and along the bottom of the diagram, therefore each repeat is present twice.. The threshold is set to show repeats greater than 1kb in length and greater than 99% identical.*

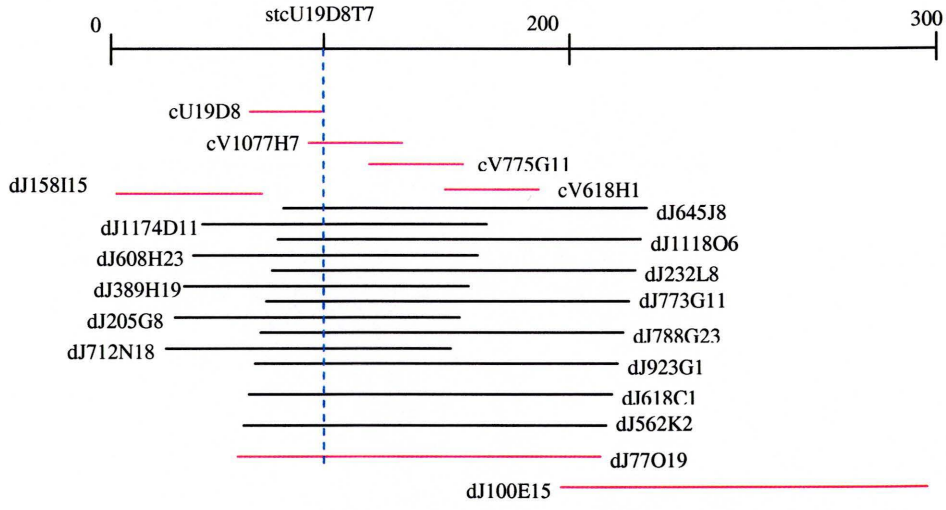
The largest repeat identified is a 140 kb inverted repeat separated by less than 10 kb (see Figure 3.14a). During the construction of the 6 Mb contig, clones mapping to this 300 kb region were thought to overlap based on fingerprinting and STS content (see Figure 3.14b). In particular an STS designed to the end of cU19D8 (stcU19D8T7) mapped to a series of PAC clones including dJ77O19. However, the sequence of dJ77O19 did not contain the sequences of cV1077H7, cV775G11 and cV618H1, but did contain a portion of sequence that matched part of the sequence of dJ158I15 and cU19D8. The regions of overlap were 100% identical but could not be a true overlap because the match was inverted. This left a gap in the contig between cV618H1 and dJ77O19 which was closed by the identification of a BAC contig constructed using stFSX2369, an X chromosome specific STS. The sequence of bA422L23 was found to contain the sequence of cV1077H7, cV775G11 and cV618H1. The sequence of bA353J17 closed the remaining gap between cV618H1 and dJ77O19 (see Figure 3.14c). An STS designed outside the region of duplication (stdJ77O19.1) confirmed the correct placement of the clones.

Figure 3.14: (see over) *Analysis of 140 kb indirect repeat (a) An enlargement of a section of Figure 3.13 showing the region of the sequence containing the repeat 4. (b) The contig as constructed before the identification of the duplicated repeat. An STS designed to the end of cU19D3 (stcU19D3T7) identified 14 clones whose overlap was confirmed by fingerprinting. The clones shown in red were identified for sequencing. (c) The final contig constructed using genomic sequence information confirmed by STS content. The clones for which genomic sequence was available are shown in red. Clones identified by the STS stFSX2369 were incorporated into the contig by sequence comparison. bA422L23 overlapped with four cosmids, cU19D8, cV1077H7, cV715H1 and cV618H1. bA353J17 overlapped bA422L23, cV618H1 and dJ77O19. Two positions for stFSX2369 were also identified. stdJ77O19.1 was designed outside the duplication and used to confirm the position of the remaining PACs.*

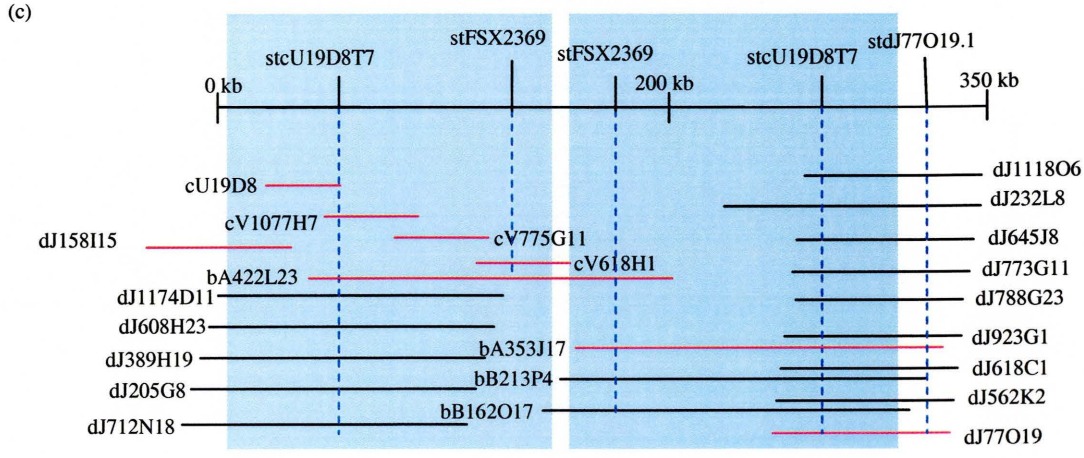
(a)



(b)


0
stcU19D8T7
200
300
cU19D8
cV1077H7
cV775G11
dJ158I15
cV618H1
dJ645J8
dJ1174D11
dJ1118O6
dJ608H23
dJ232L8
dJ389H19
dJ773G11
dJ205G8
dJ788G23
dJ712N18
dJ923G1
dJ618C1
dJ562K2
dJ77O19
dJ100E15

(c)


stcU19D8T7
stFSX2369
stFSX2369
stcU19D8T7
stdJ77O19.1
0 kb
200 kb
350 kb
cU19D8
cV1077H7
dJ158I15
cV775G11
cV618H1
bA422L23
dJ1174D11
dJ608H23
dJ389H19
bA353J17
dJ205G8
bB213P4
dJ712N18
bB162O17
dJ1118O6
dJ232L8
dJ645J8
dJ773G11
dJ788G23
dJ923G1
dJ618C1
dJ562K2
dJ77O19

3.4.4 Analysis of clone instability

The final gap to be closed in the sequence covered a region of DNA around DXS24 that had proven to be unstable in YACs (see Figure 3.15). The region containing DXS24 had previously been placed between DXS83 and DXS101C by genetic mapping. Data from the YAC map (Vetrie, D., *et al.*, 1994) gave conflicting evidence to suggest either it mapped outside the region, or that YACs containing DXS83 and DXS101C were deleted for the region containing DXS24. A single YAC was identified with DXS24, and estimated to be approximately 50 kb. The region containing DXS24 was anchored to a region distal to DXS83 by bacterial clone mapping when cU105G4, a DXS24-positive cosmid, was shown to overlap with dJ79P11. The other end of dJ79P11 was mapped to a DXS83-containing contig. This evidence agreed with both the previous placement of DXS24 by genetic mapping, and the subsequent hypothesis by Vetrie, D., *et al.* (1994) that YACs positive for DXS83 and DXS101c were deleted for DXS24.

Linking DXS24 to DXS101C proved more difficult. The clone dJ823F3 was identified using STSs designed to the ends of both cU177E8 and dJ42120 (see Figure 3.15) and selected for genomic sequencing. However the sequence of dJ823F3 revealed that although it showed overlap with a portion of dJ79P11, no common sequence was found with cU105G4. The overlap with dJ79P11 was in the same orientation and fibre-fish was carried out (by Pawandeep Dhami, data not shown) to confirm a deletion in dJ823F3 rather than a direct repeat.

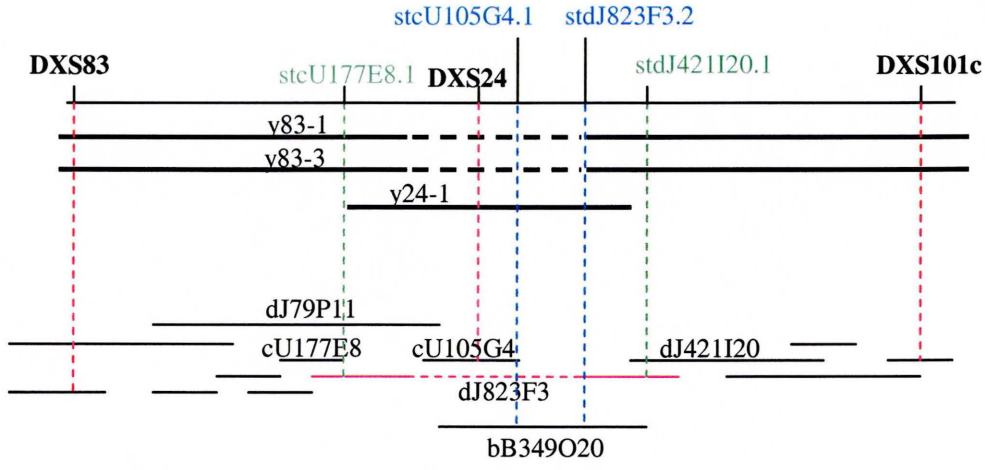


Figure 3.15: *Analysis of clone instability showing the region around DXS24 and the status of the mapping. The YAC map published by Vetrie, D., et al (1994) showed YAC clones, y83-1 and y83-2, were deleted for DXS24 (indicated by thick dotted lines), and only one YAC, y24-1, was only positive for DXS24, and no other surrounding markers. T he sequencing of dJ823F3 (shown in red), isolated with stcU177E8.1 and stdJ421I20.1 appeared to reveal a similar deletion (indicated by a thin dotted line) to that seen in the YACs. The sequencing of BAC bB349O20, identified with stcU105G4.1 and stdJ823F3.1, closed the gap in the sequence. Unlabelled thin black lines represent surrounding clones anchoring DXS24 to both DXS83 and DXS101c.*

The deletion is a similar phenomenon to that which had been seen in the YAC clones. Two further STSs (stcU105G4.1 and stdJ823F3.2) were used to identify bB349O20, which when sequenced, bridged the remaining sequence gap. Analysis of the region deleted in dJ823F3 revealed that the proximal and distal boundaries of the deletion were positioned within *Alu* repeats, which contain 12 bp of identical sequence.

3.5 Discussion

This chapter describes the construction of a 6 Mb sequence-ready bacterial clone map covering the region of Xq22 between DXS366 and DXS1230. The contig was built in four stages using the best resources available at the time for each stage (see Figure 3.16). Initial coverage across the region was gained using overlapping cosmid clones, which have insert sizes of approximately 40 kb. Although these have been used extensively in the past to map regions of the human genome and genomes of other organisms (Coulson, A., *et al.*, 1988, Doggett, N. A., *et al.*, 1995), larger insert bacterial clones are a more suitable reagent for sequencing as they enable a greater amount of sequence to be generated using a fewer number of clones. The PACs have an average insert size of 120 kb (Ioannou, P. A., *et al.*, 1994) and the improvements in cloning in the latest BAC libraries mean these have an average insert size of about 180 kb (Shizuya, H., *et al.*, 1992). Therefore the later stages of the construction of the contig reflect these improvements. Early walking between cosmid contigs identified PACs, but the later gaps were closed using BACs.

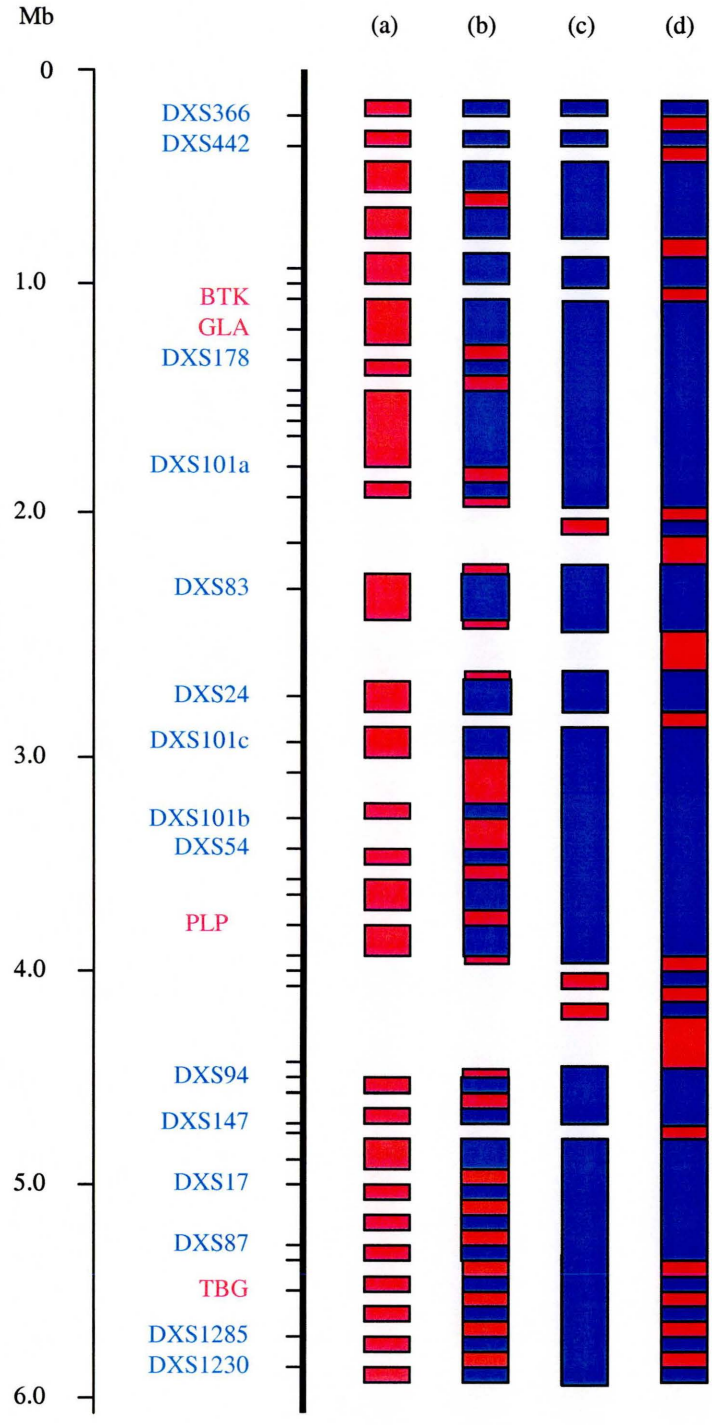


Figure 3.16: *The status of the mapping at each stage of contig construction. Bars represent the contigs; new coverage gained at each stage is shown as red bars and existing coverage is shown as blue bars. The four stages are (a) cosmid fingerprinting, (b) PAC identification using whole cosmid hybridisation, (c) seeding contigs in gaps and (d) final gap closure.*

The bacterial clone maps covering large portions of the human genome (Bentley, D. R., *et al.*, 2001; Bruls, T., *et al.*, 2001) have been constructed using the equivalent of the last two stages of this project. Initial coverage was gained using a high density of markers across a given region to identify bacterial clones and gap filling was achieved with STSs from the ends of clones at the ends of contigs. A density of markers greater than fifteen per megabase has proved sufficient to cover 80% of large regions, eg, whole chromosomes, in sequence-ready contigs (Bentley, D. R., *et al.*, 2001). If the contig described in this chapter was generated today, the start point would be the identification of BAC clones using the available STS landmarks as probes (see Figure 3.16c and d).

The integration of cosmids with PACs and BACs to form a single map had two major problems. The first is the difficulty of identifying significant overlaps between a cosmid and a larger insert clone such as a PAC or BAC using fingerprinting. FPC reports overlaps based on the probability of two clones overlapping by chance, the lower the score, the less likely the match is a random event (see Section 2.23.2). In calculating the probability score, the number of bands in common between the two clones is taken into account relative to total number of bands in each clone fingerprint. Probability of overlap is then assessed on this basis. Table 3.3 shows a set of FPC probability scores between a series of clones in the final contig.

Table 3.3: *Example of probability of overlaps, comparing clones of different sizes*

Clone 1	Number of Bands	Clone 2	Number of Bands	Matches	Probability	Actual Overlap
cV467E10	20	cU46H11	25	11	2e-05	30 kb
cV362H12	18	dJ839M11	35	10	6e-04	25 kb
dJ3E10	44	dJ197J16	55	21	1e-10	25 kb
bA269L6	39	dJ409F10	28	14	7e-04	0 kb

The results show that even though the cosmid cV362H12 and the PAC dJ839M11 overlap by 25 kb, the reported probability is the same as that reported between two non-overlapping large insert clones bA269L6 and dJ409F10. In this way, a significant overlap between a cosmid and PAC or BAC could be missed. In general a threshold of probability can be set when contigs are constructed using similarly sized clones e.g. 1e-04 for cosmids and 1e-10 for PACs/BACs.

The second major problem with constructing a contig using clones of different lengths is identifying the minimum set for sequencing. After the initial assembly of cosmid clones, a minimum set of clones from each contig was identified and sequenced. The larger insert PAC or BAC clones were added at the ends of contigs to either close gaps or extend coverage into the gaps. In most cases, when a PAC or BAC clone was sequenced, at least one cosmid at the end of each contig became redundant. A more efficient procedure for selecting a minimum set of clones for sequencing is to generate a complete contig in similarly sized clones and then choose the clones for sequencing.

The availability of the genomic sequence between DXS366 and DXS1230 allowed for a comparison of previously published maps and an assessment of the accuracy of the different types of maps. The physical maps that were compared in Section 3.3 were generated as part of the overall aim to map, sequence and analyse large regions of the human genome and have proven vital in generating the final bacterial clone contig map described in this chapter. Although differences in marker order were observed between the published maps, these can be accounted for by the limitations of each method used. Genetic mapping and radiation hybrid mapping rely on

recombination and radiation-induced DNA breakage respectively. These events do not necessarily occur randomly with respect to physical distance. The refinement of both maps and the accurate positioning of the markers on the genomic sequence is important for future study of the region. The refinement of the genetic map and the positioning of the genetic markers on the sequence are important as these markers are still being used to define critical regions for as yet uncloned diseases. The genomic sequence will also allow for the identification of new sequences that may be useful for genetic mapping such as previously unidentified dinucleotide repeats (e.g CA) that may be polymorphic in the population. The RH map contains STS generated from EST sequences and the placement of these STSs on the genomic sequence will aid the identification of the genes within the sequence.

The genomic sequence generated from clones selected from the bacterial clone contig described in this chapter provides the basis for a higher resolution analysis of the region than has previously been possible. One of the genes contained within the genomic sequence is PLP, and duplication of the region including the PLP gene is the primary cause of Pelizaeus Merchbacher Disease (PMD; Hudson, L. D., *et al.*, 1989; Trofatter, J. A., *et al.*, 1989), causing a gene dosage effect that is thought to lead to increased expression of the protein product and a disturbance of development or maintenance of myelin (Inoue, K., *et al.*, 1999). Work is currently underway with Karen Woodward (Institute of Child Health, London) to map the breakpoints of these duplications in different PMD families onto the sequence map, and to determine whether features within the sequence make the region susceptible to duplication.

A comprehensive transcript map that includes the region between DXS366 and DXS1230 is also being constructed using the available genomic sequence (Ian Barrett, The Sanger Institute). A systematic scanning of the available genomic sequence, using a combination of *de novo* gene prediction and similarity searches are placing previously known but unlocalised genes and identifying novel genes that are experimentally confirmed. A number of genetic disorders, for which the gene responsible has been localised to the region, remain uncloned and the sequence and ultimately the genes in the region will allow for the systematic screening for the genes responsible.

Chapter 4

Genome Landscape of Xq23-24

4.1 Introduction

The generation of large contiguous segments of human genome sequence allows for the systemic identification of the genes and other functional units contained within. As discussed in Section 1.3, one of the aims of studying the human genome is to identify all the genes present in the human genome which will provide the basis for furthering our understanding of the biological systems in which they are involved. There are also a large number of diseases for which the gene responsible remains uncloned. Identifying the genes within a region of interest because of the association with a particular disease enables screening for the causative mutations. For instance, a systematic search for the gene responsible for X-linked lymphoproliferative disease (XLP), previously localised to Xq25, resulted in the identification of SH2D1A, which was subsequently found to be mutated in patients with XLP (Coffey, A. J., *et al.*, 1998).

As part of the project to map and sequence the human X chromosome, the techniques developed during the construction of the sequence-ready bacterial clone contig in Xq22 (see previous chapter) were applied to extend this contig, and generate new contigs across all regions of the human X chromosome (Bentley, D. R., *et al.*, 2001). The sequencing of the X chromosome, based on these contigs, progressed rapidly in the Xq23-Xq25 region and was therefore chosen for an in depth study of the features encoded in the sequence.

RESULTS

4.2 Identification of genes

As a result of the work carried out for the X chromosome mapping and sequencing project by the X chromosome group, four contigs covering 25 Mb of Xq22-Xq25 were generated and a minimum set of clones identified and sequenced (see Figure 4.1). This provided the raw data for the analysis and gene identification across an 8 Mb region between DXS7598 and DXS7333 that encompasses the distal portion of Xq23, Xq24, and the proximal portion of Xq25. There are currently twelve contiguous segments of finished sequence covering 7 Mb and a further 600 kb of draft sequence is available.

A combination of similarity searches (BLASTX and BLASTN) and *de novo* gene prediction were carried out on finished sequence (see Figure 4.2). Prototypical and consensus repeats were masked using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html), and the remainder of the sequence was aligned using BLAST (Altschul, S. F., *et al.*, 1990) to all known cDNAs, ESTs and other sequences to identify similarities with previously known genes from human and other species. A modification of this approach was to translate the genomic sequence into all possible reading frames and compare the sequence of the translation products using BLAST with databases of known protein sequences.

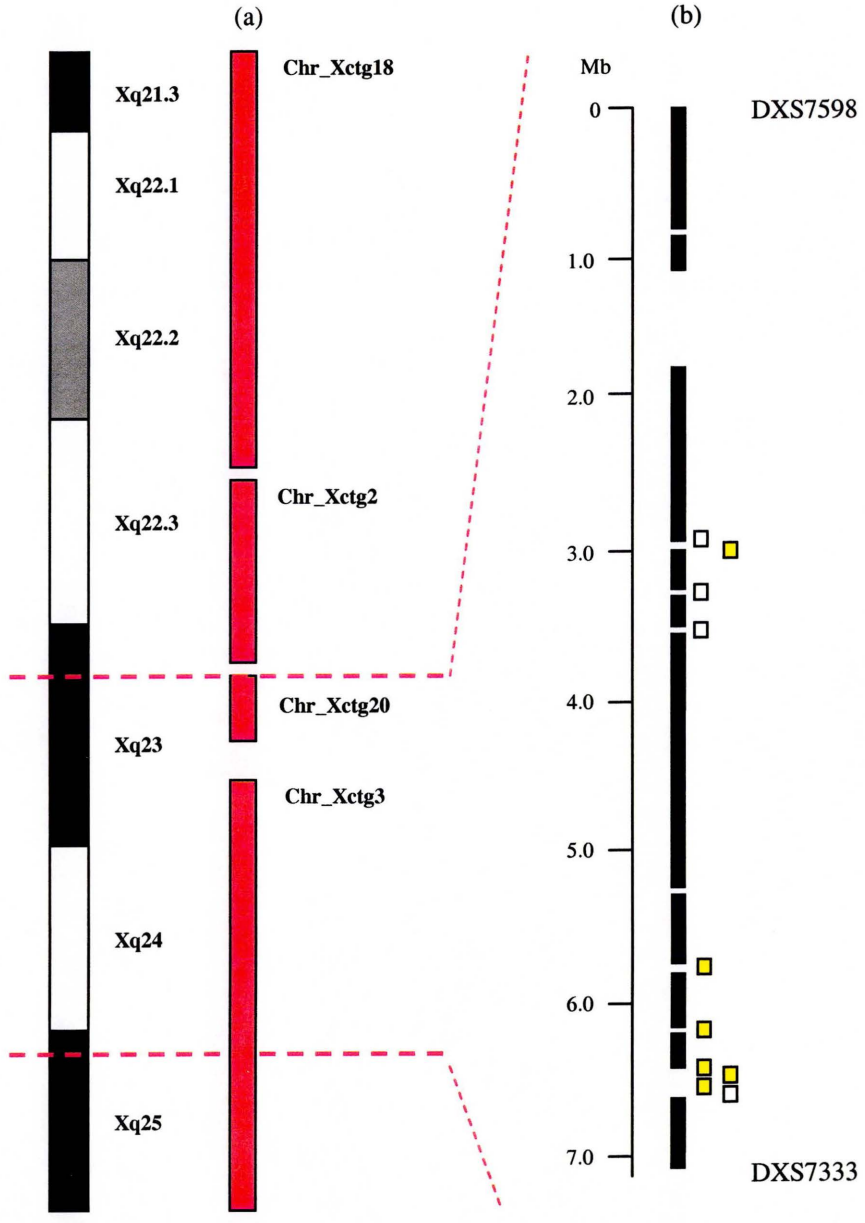


Figure 4.1: *Status of the region between Xq21.3 and Xq25 (a) Extent of bacterial clone mapping showing part of the X chromosome between Xq21.3 and Xq25, and the status of mapping and sequencing. Vertical red bars indicate contigs and their number from the Chromosome X mapping project are shown. The dotted red lines indicate the region identified for gene identification between DXS7598 and DXS7333. (b) The status of genomic sequence: black bars are continuous segments of finished sequence, yellow bars indicate clones with draft sequence available, white bars signify clones identified for sequencing, for which sequence is not yet available.*

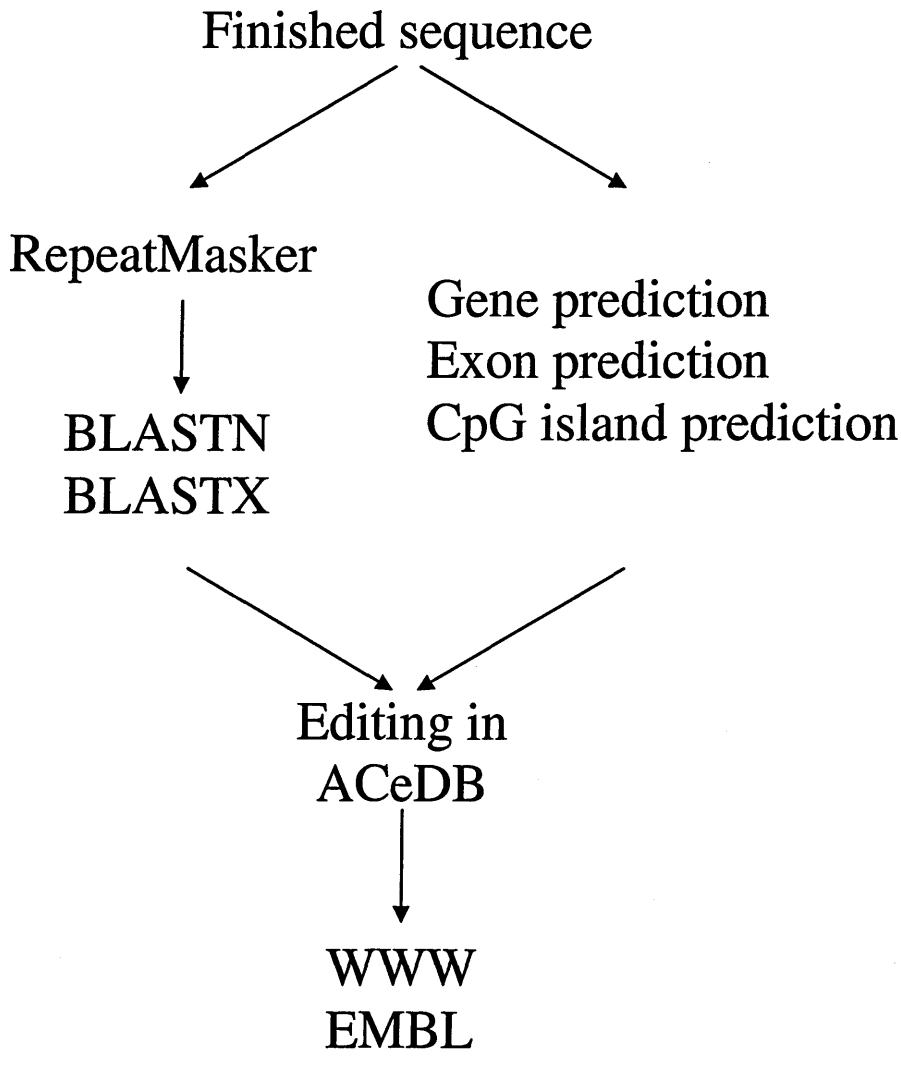


Figure 4.2: *Genomic sequence analysis. Finished sequence is analysed for repeats using RepeatMasker and compared to known protein and DNA sequences. De novo gene prediction is carried out on non-Repeatmasked sequence. The annotated sequence is viewed in an X chromosome ACeDB, Xace (see Section 2.23.3) and made available on the WWW (Sanger FTP site and EMBL).*

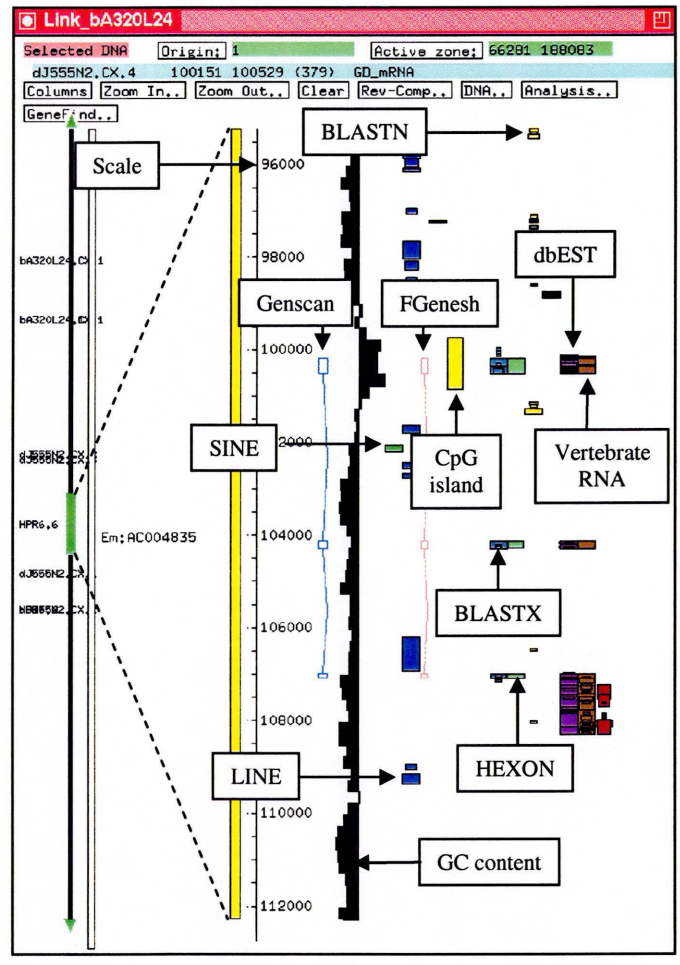
De novo gene prediction was carried out using a variety of prediction programmes to identify putative exons and genes. Also, given that CpG islands are associated with approximately 56% of genes (Hannenhalli, S., *et al.*, 2001), a CpG island finder was used (courtesy of Gos Micklen). Members of the sequence annotation group at the Sanger Centre carried out the initial annotation of the genome sequence and the results were visualised in an X chromosome specific implementation of ACeDB, Xace (see Figure 4.3).

Initial analysis of the genomic sequence identified 19 genes that were previously known (see Table 4.1). In all cases the known genes were identified because there was a full length mRNA sequence aligning exactly to the genomic sequence.

Table 4.1: *Known Genes with full length mRNA sequence*

Gene Name	Accession number	Reference
HOM-TES-85	AF124430	direct submission
hATB0+	AF151978	Sloan, J. L., *et al.*, 1999
ANT2	J02683	Battini, R., *et al.*, 1987
NDUFA1	U54993	Au, H. C., *et al.*, 1999
LAMP2	J04183	Kannan, K., *et al.*, 1996
GLUD2	X66310	Shashidharan, P., *et al.*, 1994
GRIA3	X82068	direct submission
T-plastin	M22299	Lin, C. S., *et al.*, 1993
NRF	AJ011812.2	Nourbakhsh, M., *et al.*, 2000
IL13R	X95302	Caput, D., *et al.*, 1996
ZNF-kaiso	XM_010435	direct submission
HPR6.6	Y12711	Gerdes, D., *et al.*, 1998
ZNF183	X98253	Frattini, A., *et al.*, 1997
UBE2A	M74524	Koken, M. H., *et al.*, 1996
ATP1B4	AF158383	Pestov, N. B., *et al.*, 1999
SMT3B	X99585	Lapenta, V., *et al.*, 1997
SEP2	D50918	direct submission
RPL39	U57846	Delbruck, S., *et al.*, 1997
U69a	Y11163	direct submission

(a)



(b)

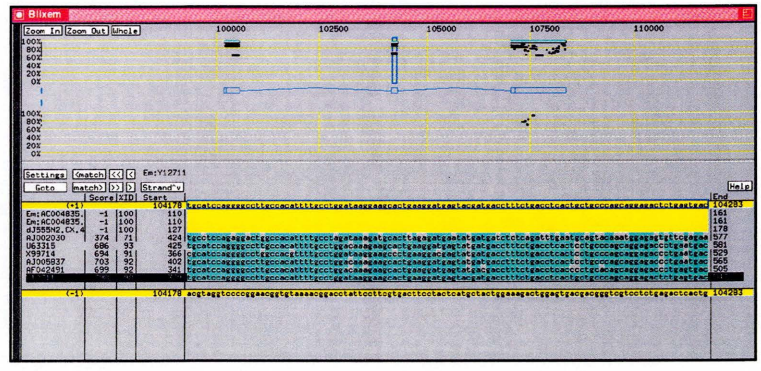
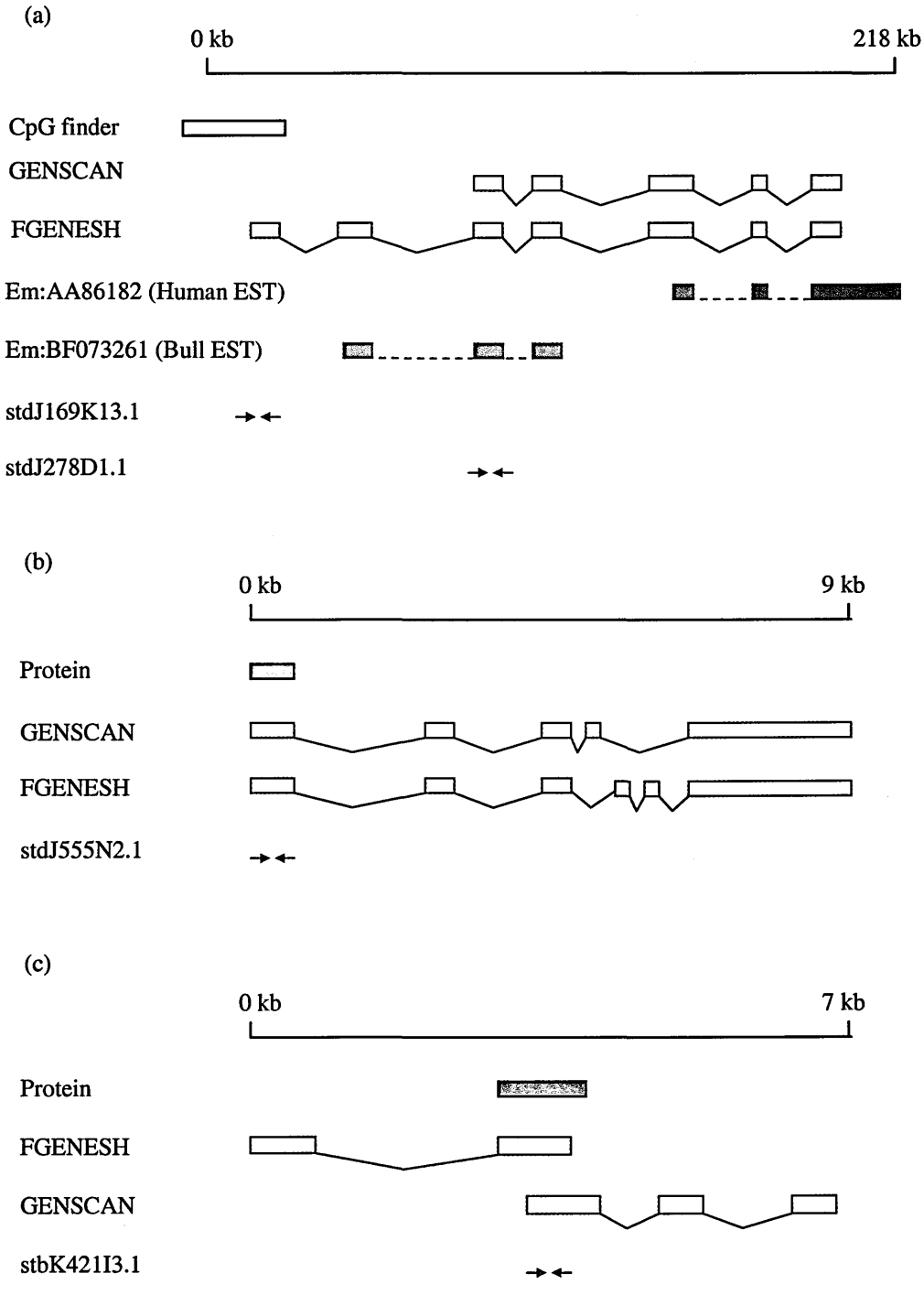


Figure 4.3: *ACeDB and BLIXEM (a) Example of annotated sequence in Xace. Features such as Genscan and Fgenesh predictions are labelled. The width of the each bar is an indication of the similarity between the genomic sequence and feature. (b) View of BLIXEM showing alignment of vertebrate mRNA sequences to the genomic sequence (see Section 2.23.4).*

Fourteen of the nineteen known genes have associated publications and the mRNAs of the remaining five genes were deposited directly into the sequence databases. The gene names are given as recommended by the Human Gene Nomenclature Committee (HGNC – http://www.gene.ucl.ac.uk/nomenclature). Although these genes needed no experimental verification, the precise exon/intron structure for each gene has been elucidated and their position and transcriptional direction on the genomic sequence in relation to neighbouring genes determined by alignment to the genomic sequence as part of this study.

In order to identify novel genes in the region, the genomic sequence was analysed for regions predicted to represent exons, based on sequence similarity searches and gene prediction programmes. Primers for the PCR were designed to regions with a variety of evidence suggesting the presence of a gene. For instance, eight pairs of primers were designed to regions predicted to be coding only by gene prediction programs, three pairs of primers were designed to regions predicted only by protein homology and two pairs of primers were designed to regions predicted only by EST homology. In some cases a protein or DNA sequence spliced across a series of exons in the genomic sequence and a predicted gene structure was identified and in other cases only a single exon was suggested. Examples of both predicted gene structures and a single exon region are shown in Figure 4.4.

Figure 4.4: *(see over) Examples of features for which STSs were designed for cDNA isolation. (a) A region of 218 kb was predicted to be coding by both GENSCAN (red boxes and lines) and FGENESH (blue boxes and lines), the 5' end suggested by the presence of a CpG island (yellow box). Two ESTs (purple boxes, splicing indicated by dotted lines) matched the genomic sequence exactly. Two STSs (indicated by red arrows) were designed, to generate novel cDNA sequence in regions not covered by the human cDNA sequence. (b) A region of 9 kb was predicted to be coding both by GENSCAN and FGENESH. A protein match (light blue box) was also observed in the first exon. (c) An example of a single exon feature where one exon from a GENSCAN prediction overlapped with one exon from a FGENESH prediction. A protein match was also observed.*


(a)
0 kb
218 kb
CpG finder
GENSCAN
FGENESH
Em:AA86182 (Human EST)
Em:BF073261 (Bull EST)
stdJ169K13.1
stdJ278D1.1
(b)
0 kb
9 kb
Protein
GENSCAN
FGENESH
stdJ555N2.1
(c)
0 kb
7 kb
Protein
FGENESH
GENSCAN
stbK421I3.1

Analysis of the genomic sequence identified twenty-two predicted gene structures and twenty single exon regions. During the experimental verification process, mRNA sequences for five genes (T-plastin, ZNF-kaiso, UPF3B, NRF and HATB0+) were published and/or submitted to sequence databases by external groups and these previously predicted genes became "known genes" and are listed in Table 4.1. In order to analyse the 42 predicted gene structures and single exon predictions, a total of 58 primer pairs were designed either automatically using PRIMER (http://www.sanger.ac.uk/cgi-bin/primer3.cgi) or manually (see Section 2.15.1). Where possible each primer was between 18 and 20 nucleotides in length and had a GC content of approximately 50%. Primers were designed within a single predicted exon and pre-screened to determine the optimal annealing temperature for the PCR. PCR was carried out on pools of up to 19 different cDNA libraries (see Section 2.8.3) to identify the individual cDNA pool likely to contain the cDNA of interest. Each cDNA library in the panel comprises approximately 500,000 clones divided into twenty-five pools, each containing 25,000 cDNA clones. Five pools were combined to form a super pool containing 100,000 cDNA clones (cDNA library resources were kindly provided by Jackie Bye). For cDNA isolation using SSPCR, primers were initially screened against the super pools and then against pools representing up to five different positive super pools. For cDNA isolation using vectorette PCR, only the super pools were screened. The results are summarised in Table 4. 2.

Table 4.2: *Experimental verification of predicted genes (see 2.8.3 for library pool codes); STSs are described in Table 2.6. Those superpools for which the equivalent pools were screened are shown in red. A gene name is given (last column) if the STS that was designed was used to generate novel cDNA sequence for confirmation of a predicted gene structure.*

STS name	Evidence	Library Pools	Positive Superpool	Positive pool	Gene
stbA45J1.1.1	EST, GENSCAN, FGENESH	V 1-11	FL:B	-	bA45J1.CX.1
stbA125M24.1.1	Protein, mRNA (human)	V 1-11	No pools +ve	-	bB125M24.1
stbK421I3.1	Protein, GENSCAN, FGENESH	S 1-17	No pools +ve	-	-
stbK421I3.2	mRNA, GENSCAN, FGENESH	S 1-17	Uact:C	Uact:14	bK421I3.CX.2
stbK421I3.3	EST, Protein, FGENESH	S 1-18	WU:E, T:ABCDE	WU:22.23, T:1.2.4.6.7.9.10	bK421I3.CX.1
stdA155F9.1	Protein	V 1-11	WH:D, DAU:BCDE	-	dA155F9.CX.1
stdA155F9.2	Protein	V 1-11	1.1.1.1 ALU:AE	-	dA155F9.CX.1
stdJ29I24.1	GENSCAN, FGENESH	S 1-18	No pools +ve	-	-
stdJ57A13.1	GENSCAN, FGENESH	S 1-17	No pools +ve	-	-
stdJ57A13.2	EST, GENSCAN	S 1-17	WH:CE, NK:CE, DAU:B	WH:14, 24	genomic contamination
stdJ93I3.1	GENSCAN, FGENESH	S 1-18	No pools +ve	-	-
stdJ93I3.2	EST	S 1-18	No pools +ve	-	T-plastin
stdJ169K13.1	FGENESH, CpG island	S 1-19	HPB:C	HPB:13.15	dJ169K13.CX.1
stdJ169K13.2	Protein, EST	S 1-17	No pools +ve	-	-
stdJ169K13.3	Protein, EST	S 1-17	2 No pools +ve	-	-
stdJ170D19.1	GENSCAN, FGENESH	S 1-18	T:E	T:21	no cDNA product
stdJ170D19.2	EST – no splice	S 1-18	No pools +ve	-	-

stdJ222H5.1	EST – no splice	S 1-17	No pools +ve	-	-
stdJ222H5.2	Protein	S 1-17	No pools +ve	-	-
stdJ278D1.1	Protein, EST, GENSCAN, FGENESH	S 1-17	HPB:ABC, SK:AB, T:E	HPB:5, SK:5, T:21.22	dJ169K13.CX.1
stdJ318C15.1	Protein, EST, GENSCAN, FGENESH	S 1-17	DAU:BCDE, HPB:BC, Uact:AD, FB:B	DAU:7.8.9, HPB:8, Uact:3.18, FB:8	dJ318C15.CX.1
stdJ321E8.1	GENSCAN, FGENESH	S 1-19	No pools +ve	-	-
stdJ321E8.2.1	EST, GENSCAN, FGENESH	V 1-11	3 T:BDE	-	dJ321E8.CX.2 dJ321E8.CX.3
stdJ321E8.3.1	EST, GENSCAN, FGENESH	V 1-11	3.1.1.1 T:E	-	dJ321E8.CX.2 dJ321E8.CX.3
stdJ327A19.1	EST, FGENESH	S 1-17	YT:ABCDE, HPB:AC, FB:ABCE, FL:BC, HL:ABCDE, SK:ABCDE, FLU:BCDE, DX3:ABCDE	YT:1.3.4.5, FB:7, SK:1.4.5, FLU:9, DXS:1.2.3.5	UPF3B
stdJ327A19.2	mRNA (mouse), EST, GENSCAN	S 1-17	No pools +ve	-	dJ327A19.CX.4
stdJ327A19.3	BLASTX, GENSCAN, FGENESH	S 1-17	YT:CD, HPB:B, FB:D, FL:C, HL:A, SK:ABC, T:E, AL:A, FLU:A	HPB:6.7, SK:2.3.4.5, FLU:5	dJ327A19.CX.3
stdJ327A19.4	BLASTX, GENSCAN, FGENESH	S 1-17	YT:ABCDE, HPB:BCE, FB:D, FL:C, HL:AE, SK:ABCD, T:CE, AL:AE, FLU:AD	YT:2.4, HPB:6.7, HL:5, SK1.5, FLU:5	dJ327A19.CX.3
stdJ327A19.5	BLASTX, GENSCAN, FGENESH	S 1-17	WP:A	WP:3	dJ327A19.CX.3
stdJ327A19.6	BLASTX, GENSCAN, FGENESH	S 1-17	WU:CE, YT:CD, DAU:D, HPB:ABCDE, FL:C, SK:ABCD, FLU:ACDE	WU:11.12, YT:13, HPB:7.8.10, SK:6.7.8.9.10, FLU:2.5	dJ327A19.CX.3
stdJ378P9.1	EST – no splice	S 1-18	DX3:ABDE, FB:ACE, FL:D, HL:E, FLU:BCDE	FB:4, FLU:4.6	genomic contamination
stdJ394H4.1	GENSCAN	S 1-18	No pools +ve	-	-
stdJ404F18.1	EST, FGENESH	S 1-18	WU:AB, NK:ABDE, HPB:BE, BM:B, HL:A, FLU:A, AL:E	NK:3, HPB:10, BM:7	dJ1139I1.CX.1
stdJ404F18.2	EST, GENSCAN, FGENESH	S 1-17	WU:BC, YT:CD, NK:C, DAU:E, HPB:ABDE, BM:ACE, FB:CE, SK:BDE, FLU:E, DX3:C	YT:14, NK:15, SK:6.8	dJ876A24.CX.1

stdJ404F18.3	EST, GENSCAN, FGENESH	S 1-18	WU:BCE, YT:BCDE, HPB:ABDE, Uact:ABDE, DX3:AC, FB:CDE, HL:B, SK:BCDE, FLU:BE, ALU:B, AH:CE	ALU:6.8.10	dJ876A24.CX.1
stdJ452H17.1	mRNA (mouse), EST	S 1-18	No pools +ve	-	-
stdJ452H17.1.1	mRNA (mouse), EST	S 1-19	T:CD	T:13	no cDNA product
stdJ525N14.1	Protein, GENSCAN, FGENESH	S 1-17	WU:E, T:ABCDE	WU:22, T:21.25	dJ525N14.CX.1
stdJ525N14.2	Protein, EST	S 1-17	NK:C	NK:15	Genomic contamination
stdJ525N14.3	FGENESH	S 1-17	No pools +ve	-	-
stdJ525N14.4	mRNA (mouse)	S 1-17	WU:ACE, DAU:ABCD, HPB:AB, BM:A, Uact:CE, SK:A, FLU:CD	WU:5, DAU:4.5, Uact:1.3, SK:4	ZNF-kaiso
stdJ525N14.5	mRNA (mouse)	S 1-17	NK:ABCDE, DAU:ABCDE, BM:ACDE, Uact:D, FB:BE, HL:E, T:E, AL:C	NK:1.2.5, DAU:4.5, BM:1, FB:6, T:5	ZNF-kaiso
stdJ525N14.6	mRNA (mouse)	S 1-18	All pools +ve	Stopped due to poor primer design	ZNF-kaiso
stdJ525N14.7	mRNA (mouse)	S 1-18	WU:CE, WH:DE, DAU:ABD, HPB:B, SK:AE, FLU:CD	DAU:4, HPB:9, FLU:14, SK:4, WH:16, WU:12.15	ZNF-kaiso
stdJ525N14.10	Protein	S 1-18	No pools +ve	-	
stdJ555N2.1	Protein, GENSCAN, FGENESH	S 1-18	No pools +ve	-	dJ555N2.CX.1
stdJ562J12.1	Protein	S 1-19	FB:ABC	FB:1	dJ562J12.CX.1
stdJ655L22.1.1	mRNA (human)	V 1-11	WU:ADE, FB:ABCD, FL:ABD, FLU:C, HL:C, ALU:AE, T:ABCDE, SK:ABCDE	-	dJ655L22.CX.1
stdJ755D9.4	Protein	S 1-18	No pools +ve	-	-
stdJ808P6.1	Protein	S 1-18	FB:ADE, FL:ABCE, HL:ACE, SK:A	stopped	HATB0+
stdJ808P6.2	Protein	S 1-18	WH:BC, YT:C, NK:A, DAU:DE, Uact:C, FB:AC	Stopped	HATB0+
stdJ808P6.3	Protein	S 1-18	HPB:CD, DX3:A	stopped	HATB0+
stdJ876A24.1	EST	S 1-18	WU:CE, DAU:ABCE, HPB:BCDE, BM:AE, Uact:ABCDE, FB:ABCD	DAU:2.3, HPB:6, BM:5, Uact:5	NRF

stdJ876A24.2	Protein, mRNA (human)	S 1-18	WU:ABCDE, YT:ABCDE, NK:ABCDE, DAU:ABCE, HPB:ABCDE, BM:ABCDE, Uact:ACE, DXS3:ABCDE, FB:ABCD, HL:CE, SK:ABCE, T:ABCE, FLU:ABCDE, AH:ABCDE	BM:1.2.3.5, T:1.3.5, AH:2.4.5, FB:3	Septin2
stdJ876A24.4	EST	S 1-18	YT:ABD, :DAU:BE, HPB:ABCDE, HL:B, SK:BCDE, FLU:C	T:3, HPB:1, SK:8.9, Dau:8, FLU:11	NRF
stdJ878I13.1	GENSCAN, FGENESH	S 1-18	No pools +ve	-	-
stdJ1139I1.2	FGENESH, CpG island	S 1-18	No pools +ve	-	-
stdJ1152D16.1	EST, GENSCAN, FGENESH	S 1-18	WU:BD, YT:BCDE, DAU:AE, HPB:ABDE, BM:E, DX3:C, FB:E, SK:BCDE, FLU:BE	HPB:4.5, BM:21, FB:21, SK:8.10	dJ876A24.CX.1

Thirty-six of the 58 primer pairs screened gave positive superpools in the libraries tested. Analysis of the twenty-two that failed to give positive superpools showed that eight were designed to regions predicted to be coding by gene prediction programs alone. The remaining fourteen were predicted by a combination of protein matches, EST matches and gene prediction program. Twenty-six of the thirty-six primer pairs were screened against the cDNA library pools for cDNA isolation by SSPCR and as expected all gave positive pools. Ten STSs gave positive superpools but were not subsequently screened against the pools, because six of the ten were to be used for cDNA isolation using vectorette PCR and a the remaining four were stopped because a mRNA was deposited into the sequence databases for the hATB0+ gene making cDNA isolation unnecessary.

cDNA isolation from individual positive pools was carried out using either SSPCR (Huang, see Figure 4.5) or vectorette PCR (adapted from Riley, J., *et al.* (1990), see Figure 4.6) (see also Section 2.22.3 (Figure 2.1) and 2.22.4 (Figure 2.2) for schemas). For each predicted gene, cDNA isolation was carried out on three pools or super pools from different cDNA libraries in order to increase the likelihood of generating a cDNA sequence covering the entire predicted gene. When different sized products were generated in different pools, cDNA products were chosen for sequencing based on length (where possible the largest band was sequenced), but also intensity (the strongest band took precedence over the largest band). All products generated for sequencing were assigned an Sanger Centre cDNA number (sccd) prior to sequencing.

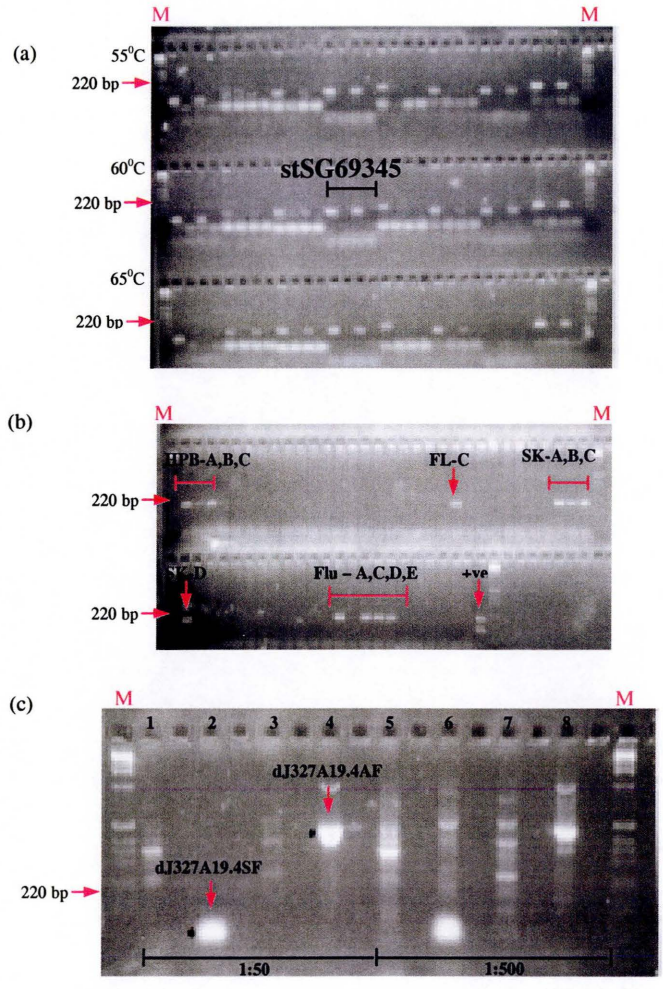


Figure 4.5: *cDNA isolation by SSPCR (a) Eight STSs designed to exons of predicted genes were tested for the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures (M = marker). (b) One of the STSs, stSG69345, was used to amplify DNA of pools of cDNA clones from 14 different libraries (see Section 2.9). (c) The results of the second round of SSPCR protocol (see 2.22.1 for schema). A combination of nested sequence-specific primers (stSG77080S and stSG77080A) and vector-specific primers (1RP and 2FP) were used to amplify the products from the first round of SSPCR. Two different dilutions of template were used (1:50, lanes 1-4, 1:500, lanes 5-8). Bands from lanes two and four were excised for sequencing.*

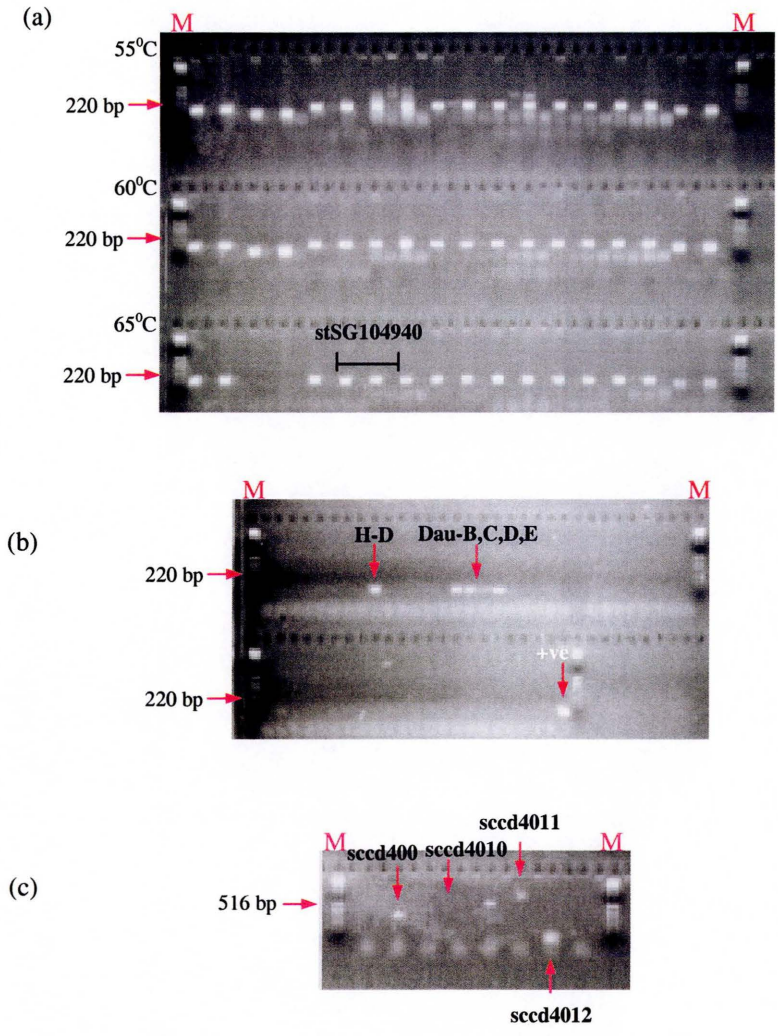
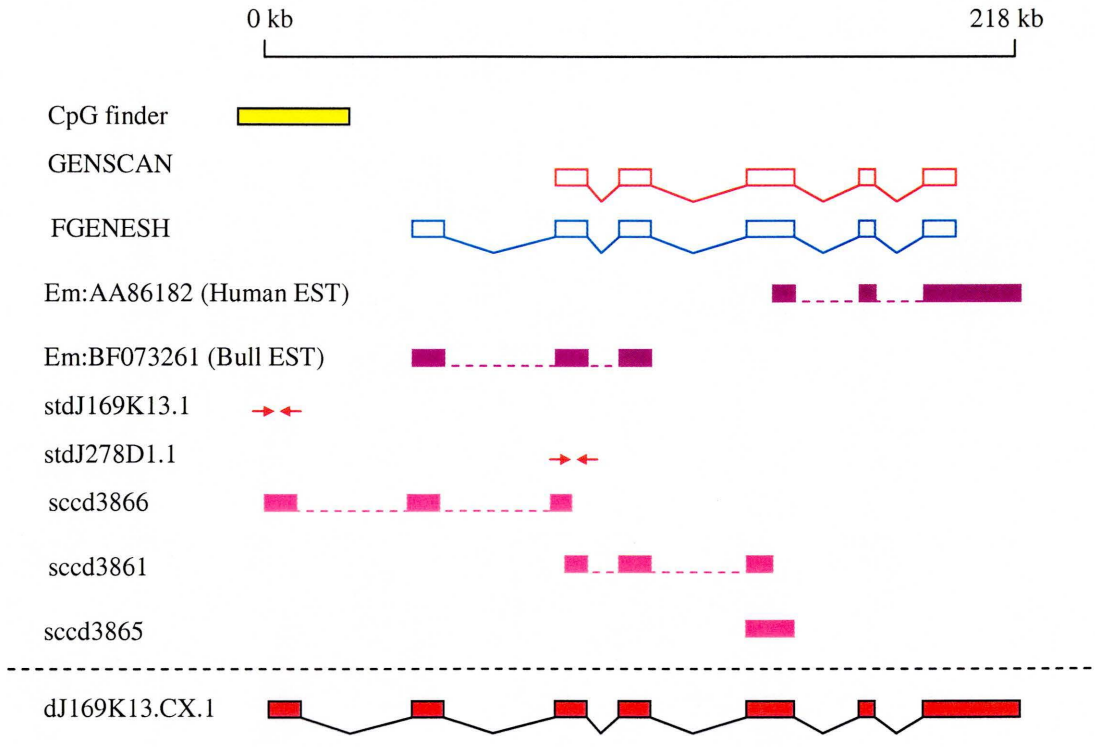


Figure 4.6: *cDNA isolation by vectorette PCR (a) Nine STSs designed to exons of predicted genes were tested for the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures of the PCR (M = marker). (b) One of the STSs, stSG104940, was used to amplify DNA of pools of 100,000 cDNA clones from 11 different vectorette libraries. (c) Results of vectorette PCR. A combination of sequence-specific primers (stSG104940S and stSG104940A) and vectorette primer 224 was used to amplify DNA of two superpools (H-D and DauB – see Section 2.9 for library informatation) at two different concentrations (1:100 and 1:1000). Bands from lanes 2, 5, 6 and 7 were excised, purified, and sequenced.*

cDNA sequence was aligned to the genomic sequence and the gene structure evaluated for possible extension. Confirmation of a predicted gene was considered complete when there was human cDNA sequence covering at least the predicted protein-coding region, and as much untranslated region (UTR) as possible. A total of fourteen predicted genes were confirmed and eleven gene structures remain unconfirmed. An example of the confirmation of one gene is shown in Figure 4.7.

Figure 4.7: *(see over) Confirmation of a novel gene. (a) dJ169K13.CX.1 (exons shown as red boxes, introns as black lines) was predicted by GENSCAN (exons shown as open red boxes linked by red lines) and FGENESH (exons shown as open blue boxes linked by blue lines), and two ESTs (shown in purple), one human and one from bull. A CpG island upstream of the predicted genes suggested a possible location for the 5' end of the gene. Three cDNA sequences (shown as pink boxes) were generated to confirm the 5' end of this gene. (b) The cDNA sequence for sccd3866. Sequence corresponding to exons as they appear in the genomic sequence are coloured as alternating red and blue open boxes.*

(a)



(b)

```
  1 gtgctctaaagctttagagaagtggtc
 31 ggggcgagcagagggtgcgaaggtgcgggt
 61 gctggtgcctcgcagcaggagggagccccg
 91 gctgcgccgcgcgactccctctttggccct
121 cggagcgcagcacccggcggacaagcggcg
151 ggacgccaggacgcggcgagcaagatctct
181 cgtggaagaggaagaccaacacatgaaatt
211 gtcccttggaggcagcgaaatgggcctctc
241 atcccatttgcagtcttccaaggcaggacc
271 tacacgcatctttaccagcaataccacag
301 ttctgtggtgttacagggctttgaccagct
331 tcgacttgaaggattgctttgtgatgtgac
361 cctgatgccaggtgacacagatgatgcttt
391 ccctgtgt
```

Twenty pseudogenes were identified within the region and are predicted to have arisen due to the reverse transcription of mRNAs into the genomic sequence. They all appear to have a functional counterpart elsewhere in the human genome. The pseudogenes were identified because they have no introns, a poly A tail within the genomic sequence and a disrupted ORF (see Figure 4.8) (see appendix, Table 4.6 for a full list of the pseudogenes identified in the region).

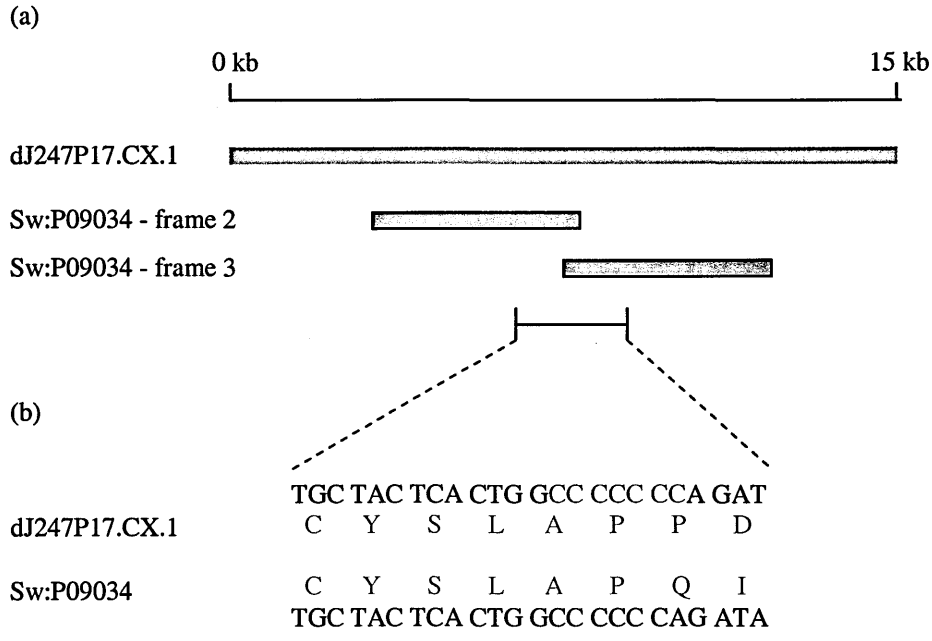
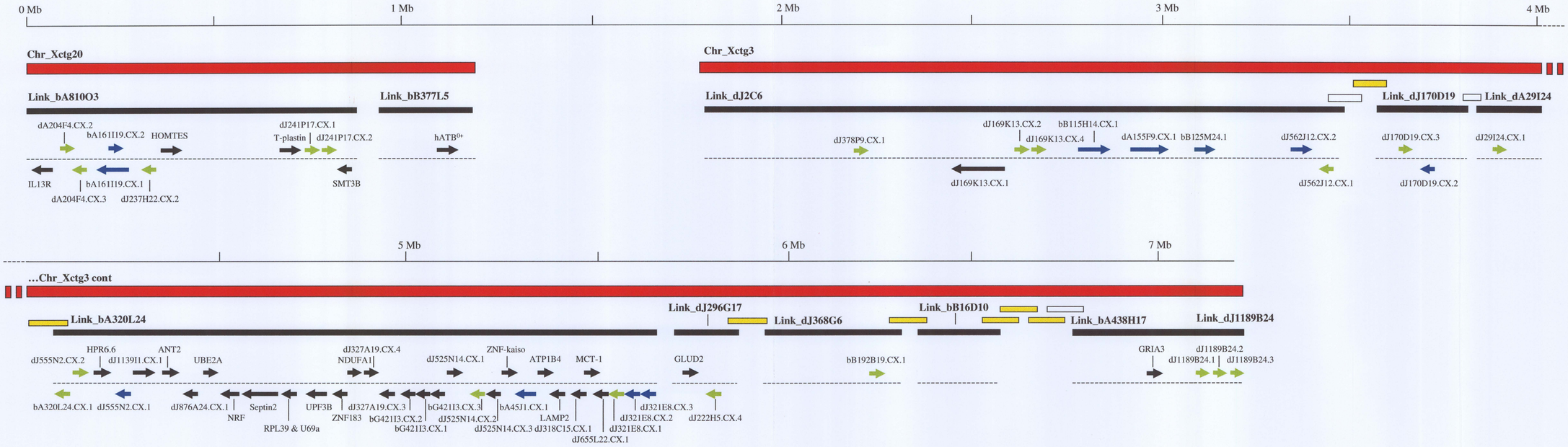


Figure 4.8: *Example of a pseudogene. (a) The extent of the dJ241P17.CX.1 (shown as a green box) is shown and is a pseudogene of arginosuccinate synthetase (ASS). Part of the protein sequence of ASS (Sw:P09034) aligns in two blocks (shown as blue boxes) (b) Disruption of the reading frame due to an insertion of a C within a run of seven C's (shown as red letters). The alignment of the nucleotide sequence (black letters) and the amino acid sequence (blue letters, using the one letter code) surrounding the insertion are shown.*

In summary, a gene map encompassing the distal portion of Xq23, Xq24 and the proximal portion of Xq25 between DXS7598 and DXS7333 covering 8 Mb has been constructed (see Figure 4.9). The region contains 33 confirmed genes (of which 14 were confirmed during this study), 11 predicted genes and 20 pseudogenes.

Figure 4.9: *(see over) A summary of the gene map between DXS7598 and DXS7333. The red bars indicate the contigs status and the black bars indicate the extent of finished sequence. Each link represents a series of individual clones (see appendix to this chapter). Yellow bars indicate clones for which draft sequence is available, and white bars indicate clones selected for sequencing, but not sequenced as of September 2001. A scale is given in megabase pairs (Mb). Approved names are given for known genes (see Table 4.1). Genes are indicated by arrows (black – complete, blue – predicted, green – pseudogene), the direction of each arrow reflects the direction of transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.*


0 Mb
1 Mb
2 Mb
3 Mb
4 Mb
Chr_Xctg20
Link_bA810O3
Link_bB377L5
dA204F4.CX.2
bA161I19.CX.2
HOMTES
dJ241P17.CX.1
T-plastin
dJ241P17.CX.2
hATB0+
IL13R
bA161I19.CX.1
dA204F4.CX.3
dJ237H22.CX.2
SMT3B
Chr_Xctg3
Link_dJ2C6
Link_dJ170D19
Link_dA29I24
dJ169K13.CX.2
bB115H14.CX.1
dJ378P9.CX.1
dJ169K13.CX.4
dA155F9.CX.1
bB125M24.1
dJ562J12.CX.2
dJ170D19.CX.3
dJ29I24.CX.1
dJ169K13.CX.1
dJ562J12.CX.1
dJ170D19.CX.2
5 Mb
6 Mb
7 Mb
…Chr_Xctg3 cont
Link_bA320L24
Link_dJ296G17
Link_dJ368G6
Link_bB16D10
Link_bA438H17
Link_dJ1189B24
HPR6.6
ANT2
dJ327A19.CX.4
ZNF-kaiso
dJ555N2.CX.2
dJ1139I1.CX.1
UBE2A
NDUFA1
dJ525N14.CX.1
ATP1B4
MCT-1
GLUD2
bB192B19.CX.1
GRIA3
dJ1189B24.2
dJ1189B24.1
dJ1189B24.3
bA320L24.CX.1
dJ555N2.CX.1
dJ876A24.CX.1
NRF
Septin2
RPL39 & U69a
UPF3B
ZNF183
dJ327A19.CX.3
bG421I3.CX.2
bG421I3.CX.1
dJ525N14.CX.2
dJ525N14.CX.3
bG421I3.CX.3
bA45J1.CX.1
LAMP2
dJ318C15.CX.1
dJ655L22.CX.1
dJ321E8.CX.1
dJ321E8.CX.2
dJ321E8.CX.3
dJ222H5.CX.4

4.3 Evaluation of genes in region

Genes can be evaluated for "completeness" by analysing the genomic sequence for the functional signals (indicated by *italics*). In order for a gene to be transcribed, an RNA polymerase binds within the *core promoter sequence* and initiates transcription at a specific position in the genomic sequence, known as the *transcription start site.* As discussed in Section 1.3, the 5' end of approximately 56% of genes lies within a *CpG island* (Antequera, F., *et al.*, 1993). Transcription proceeds along the DNA sequence until the RNA polymerase encounters a *polyadenylation signal* (consensus: AATAAA) and transcription terminates soon after. The pre-mRNA then undergoes a series of processing steps including the addition of a 3' polyA tail and a 5' cap (an additional guanine added at the 5' end), and the splicing out of the introns. Three important splice signals (*5' splice site*, *3' splice site* and the *branch point*) are involved in the removal of introns and these are recognised by the spliceosome. Approximatley 99.9% of splice sites studied conform to consensus sequences, GT at the 5' end of the intron (spliced donor), and AG at the 3' end of the intron (splice acceptor) (Levine, A., *et al.*, 2001).

Ribosomes scan the processed mRNA for a *translation start site* (usually ATG, coding for methionine). Sequence in the mRNA preceding the translation start site is termed *5' untranslated region* (5' UTR). Translation continues until the ribosome encounters a *stop codon*, a run of three bases in frame that do not code for an amino acid (UGA, UGG and UAA). Sequence following the stop codon is termed *3' untranslated region* (3'UTR).

However, because of the difficulty in identifying some of these signals, the presence or absence of such sequences within any one gene can only be used as a guide of the completeness. Furthermore, some of the signals are not found in all genes. At the 5' end of a gene, the transcription is initiated downstream of core promoter sequences such as tata boxes. Programmes such as PromoterInspector (Scherf, M., *et al.*, 2000) and Eponine (courtesy of Thomas Down) attempt to predict regions likely to contain such sequences. The optimal context for initiation of translation is GCCACCatgG (Kozak, M., 1991), but within this motif, two bases exert the strongest effect, a G at the first base after the ATG, and a purine (preferably A) three nucleotides upstream.

Confidence that the true 3' ends of genes have been identified, can be increased by identifying the polyadenylation signal in either the genomic sequence or the cDNA sequence, of which the most common is AATAAA, but there are other less common sequences used such as TATAAA (Beaudoing, E., *et al.*, 2000). In some cases aberrant cDNA clones can arise due to the priming of the poly dT primer from a polyA tract in contaminating genomic sequence, but these can be distinguished from real expressed clones as these contain a polyA tail in the cDNA sequence that is not present in the genomic sequence.

Genes within the 8 Mb region have been analysed for the presence of these features. The results are summarised in Table 4.2. All splice sites in the genes identified in this study conformed to the consensus splice site sequence - GT at the 5' end and AG at the 3' end of each intron.

Table 4.3: *Evaluation of the gene structures. PI = PromoterInspector, At least one = 5' end predicted by at least one method. A.S = Alternative Splice*

	5' end				3' end	
Gene Name	**CpG island**	**PI**	**Eponine**	**At least one**	**Poly A Signal**	**A.S**
IL13R					AATAAA	2
bA161I19.CX.1	YES		YES	YES		
bA161I19.CX.2	YES			YES		
HOM-TES-85					AATAAA	
T-plastin					AATAAA	
SMT3B		-9				
hATB0+						
dJ169K13.CX.1	YES	YES	YES	YES	AATAAA	
bB115H14.CX.1	YES	YES	YES	YES	AATAAA	
dA155F9.CX.1	YES	YES	YES	YES		
bB125M24.1						
dJ562J12.CX.2					GATAAA	
dJ170D19.CX.2					AATAAA	
dJ555N2.CX.1			YES	YES		
HPR6.6	YES	YES	YES	YES	AATAAA	
dJ1139I1.CX.1	YES		YES	YES	GATAAA	
ANT2	YES	YES	YES	YES	AATAAA	
dJ876A24.CX.1					ATTAAA	
SEP2	YES	YES	YES	YES	ATTAAA	
NRF	YES		YES	YES	AATAAA	2
UBE2A	YES	YES	YES	YES		
RPL39	YES				AATAAA	
U69a						
UPF3b	YES		YES	YES	AATAAA	
ZNF183	YES	YES		YES	AATAAA	

dJ327A19.CX.3	YES			YES	AATAAA	
dJ327A19.CX.4		-8				
NDUFA1	YES	YES		YES	AATAAA	
bG421I3.CX.1			-1		AATAAA	
bG421I3.CX.2	YES			YES		
bG421I3.CX.3			YES	YES	AATAAA	
dJ525N14.CX.1		-4	YES	YES	AATAAA	
dJ525N14.CX.3		-6	-6			
ZNF-kaiso	YES	YES		YES	ATTAAA	
bA45J1.CX.1	YES			YES	AATAAA	2
ATP1B4						
dJ318C15.CX.1		-2	-2		AATAAA	
LAMP2	YES	YES		YES	TATAAA	2
dJ655L22.CX.1	YES	YES	YES	YES	ATTAAA	3
MCT-1					AATAAA	
dJ321E8.CX.2	YES			YES		
dJ321E8.CX.3					TATAAA	
GLUD2	YES			YES		
GRIA3	YES			YES		2
Total (%)	**55**	**30**	**34**	**60**	**64**	**14**

4.3.1 Evaluation of the 5' ends

CpG rich sequences (identified by CpGfinder) are present at the 5' end of 25 of the 44 genes (56%). PromoterInspector predicts regions likely to contain promoter sequences overlapping the 5' ends of 13 genes (30 %). Eponine predicts transcription start sites at the 5' end of 15 genes (34%). The 5' end of 26 genes (60%) lies within a region predicted by at least one of the programmes. An example of the analysis for the ANT2 gene is shown in Figure 4.10a. In general, PromoterInspector and Eponine predict sequences for those genes associated with a CpG island. *Promoter* sequences or transcription start sites are predicted within 10 kb of a further 5 genes (11%). In general, both PromoterInspector and Eponine predict the presence of promoter sequences or transcription start sites at the 5' end of genes for which a CpG island has been detected. This is consistent with the analysis of the genes on chromosome 22 (John Collins, personal communication).

4.3.2 Evaluation of the 3' ends

As discussed in the previous section some of the cDNA libraries used both in this study and for other cDNA sequencing projects, have been generated by priming from the poly-A tail of mRNAs. In general, the true 3' end of an mRNA is represented in cDNA or EST sequence. However, the presence of a poly-A sequence within the mRNA, can lead to artefacts that represent sequences truncated at the 3' end in cDNAs that are generated using the poly(dT) primer. An example is shown in Figure 4.10b, where a cluster of apparently 3' EST sequences are located upstream of a second cluster of 3' EST sequences within the 3'UTR of Septin2. The presence of a

polyadenylation signal within the sequences of the second cluster increases the likelyhood that the complete 3'UTR has been identified. Analysis of all 44 genes shows that the most common polyadenylation signal (AATAAA), is present within 30 bp of the end of the transcribed sequence in 18 genes (42%). A further 10 genes (23%) contain one of three less common sequences (ATTAAA, GATAAA and AATATA) within 30 bp from the end of the transcribed sequence. These figures are slightly different to those reported by Beaudoing, E., *et al* (2000) who compared 8700 mRNAs and showed that 58% contained AATAA and 20% contained the three less common variants. It is not clear whether the genes with no detectable polyA signal represent incomplete structures, although it is possible that the genes contain other variations of the polyA signal (Beaudoing, E., *et al.*, 2000).

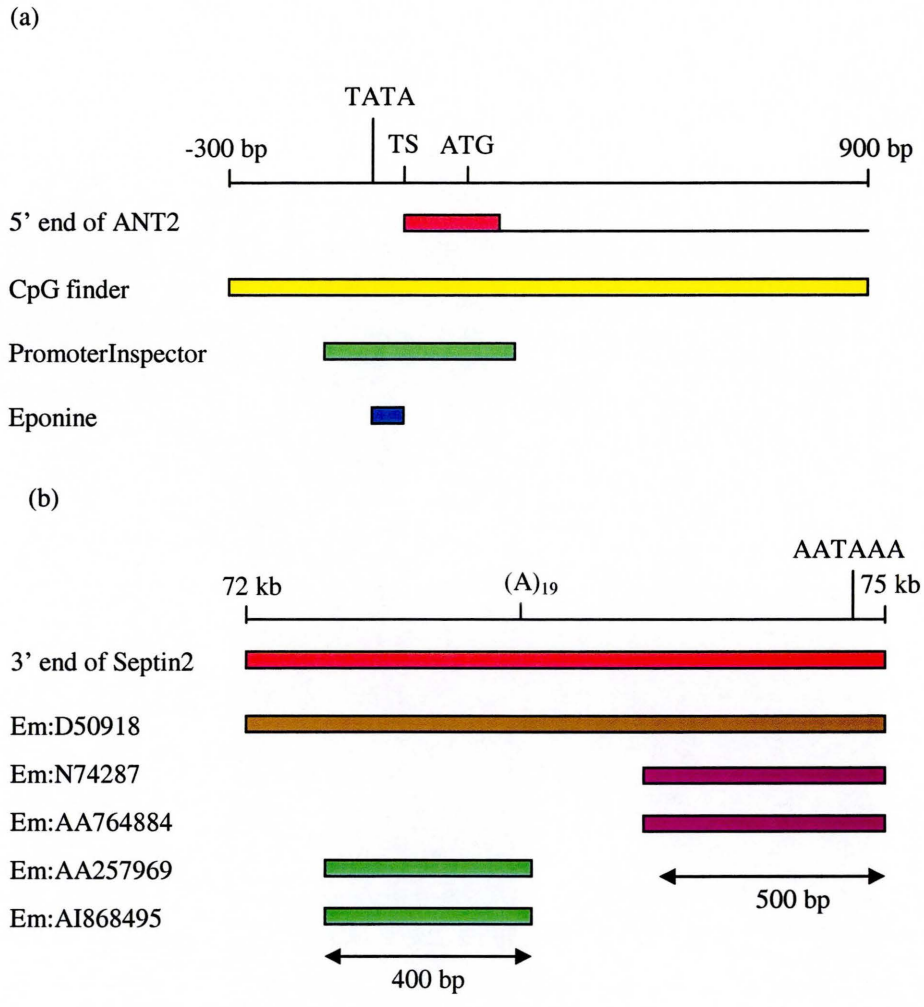


Figure 4.10: *Evaluation of gene structures (a) The 5' end of ANT2 (shown as a red box) coincides with the results of analysis using three prediction packages, CpGfinder (yellow box), PromoterInspector (green box) and Eponine (blue box) TATA = TATA box, TS = transcription start, ATG = translation start. (b) The 3' end of Septin2 gene (shown in red) aligns with the 3' end of the cDNA KIA01228 (Em:D50918, shown as a brown box), along with two EST sequences (shown as purple boxes) is the evidence for the correct 3' end of the gene. Two other EST sequences (shown as green boxes) indicate the position of a less likely 3' end on the basis that no polyA signal is observed. The correct 3' end is confirmed by the presence of poly-adenylation signal 15 bases upstream (AATAAA).*

4.3.2 Alternative splicing

As was recently stated in the publication of the draft sequence of the human genome and associated publications (IHGSC, 2001), the number of genes in the human genome is expected to be between 30,000 and 40,000. However, the generation of alternatively spliced transcripts for individual genes increases the complexity of the transcriptome in higher eukaryotes without the need to increase the number of gene loci. Analysis of the 727 genes identified on Chromosome 20 showed there was evidence for alternative splicing for 29% of the genes, with an average of 1.65 transcripts per gene (exluding putatively predicted genes) (Deloukas, P., *et al.*, 1998). In this project described in this chapter, the primary aim was to identify as many genes as possible and at least one full-length transcript. However, where evidence of alternative splicing was observed, the partial transcripts were annotated but not confirmed. In the region studied, there is evidence for alternative splicing in six genes and the average number of transcripts per gene is predicted to be 1.16, which is lower than that observed on Chromosome 20.

An example of a gene with evidence of alternative transcripts is shown in Figure 4.11a. In the example shown, three different transcripts have been identified based on three separate cDNA sequences. The first transcript, dJ655L22.CX.1 aligns with a human cDNA sequence and contains three exons. cDNA sequence generated during this project, showed splicing of exon one to exon three of dJ655L22.CX.1, and this transcript has been termed dJ655L22.CX.1b. A third transcript, dJ655L22.CX.1c shows exon one of dJ655L22.CX.1 splicing onto a novel exon, which in turn splices onto exon three of dJ655L22.CX.1. The presence of this novel exon in

dJ655L22.CX.1c is based on an EST sequence from *Bos taurus* and has not so far been seen in human cDNA sequence. A fifth exon may be present in the region, as predicted by GENSCAN, but has not been verified to date.

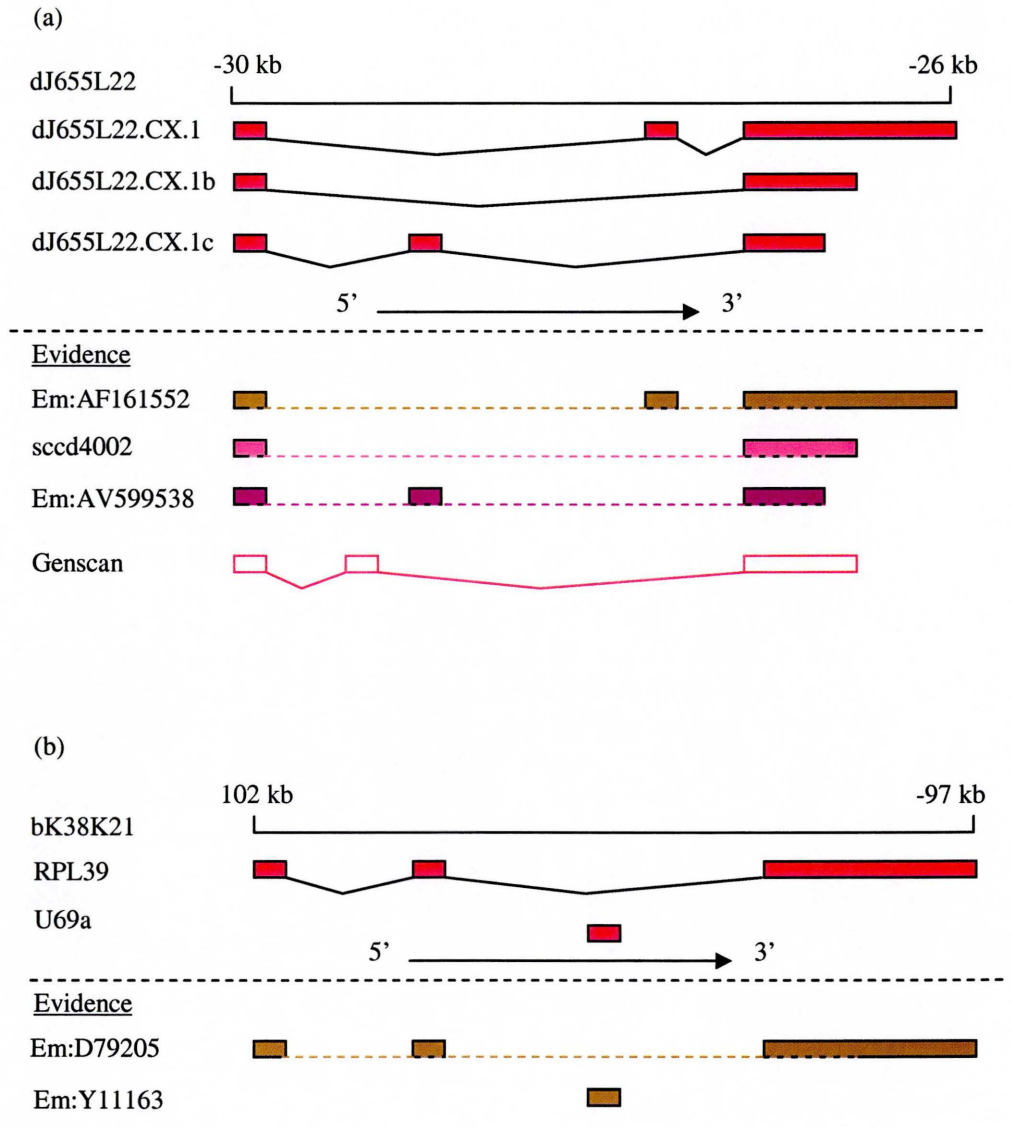


Figure 4.11: *Genes in their genomic context (1) (a) Three alternatively spliced transcripts of dJ655L22.CX.1 (shown as red boxes linked by black lines). The evidence for each transcript is shown below the dotted line. Aligned sequence is indicated by boxes linked by dotted lines. (b) The RPL39 gene (shown as red boxes linked by black lines) contains the snoRNA U69a (shown as a single red box) within intron 2. The alignment of each mRNA to the genomic sequence is shown as brown boxes below the dotted line.*

4.3.4 Genes in their genomic context

One of the advantages of systematic gene identification across large regions of sequence is the placement of genes in their genomic context. One phenomenon is the presence of a gene within an intron of a second gene. This has previously been shown in humans and other species, for instance on the human X chromosome, the F8A gene is located within an intron of the Factor VIII gene (Naylor, J. A., *et al.*, 1995). Figure 4.11b shows a similar example, where U69a, a gene encoding a small nucleolar (sno) RNA is located within intron 3 of RPL39, a ribosomal protein subunit. The positioning of these genes may play a significant role in their function as both are involved in the transcriptional machinery of the cell (Delbruck, S., *et al.*, 1997; Eliceiri, G. L., 1999).

A second example of gene placement is shown in Figure 4.12a. ZNF183 and NDUFA1, two well-characterised genes are placed on opposing strands, and their transcription start sites are predicted to be only 12 bp apart. Further investigation is needed to identify their respective regulatory elements to determine how these genes are transcribed and controlled, and whether their close proximity plays an important role in their correct functioning.

During the identification of one gene, dJ876A24.CX.1, an EST was aligned to the genomic sequence of dJ876A24 but was assigned to chromosome 8 by the Gene map '98 project (Deloukas, P., *et al.*, 1998, updated electronically 1999, see http://www.ncbi.nih.nlm.gov/genemap99) (see Figure 4.12b and c). Further analysis of the two regions, revealed a pseudogene of dJ876A24.CX.1 on chromosome 8. Primers designed within the EST for RH mapping, flanked an intron within dJ876A24.CX.1 and gave an expected genomic PCR product of 14.8 kb. This would be an X-specific product; none was observed during the RH mapping experiment. By contrast, the absence of the intron between the two primer sequences in the pseudogene on chromosome 8 resulted in the generation of a 287 bp chromosome 8-specific product. This product was detected during the original RH mapping experiments, resulting in the assignment of the EST AA085642 to chromosome 8.

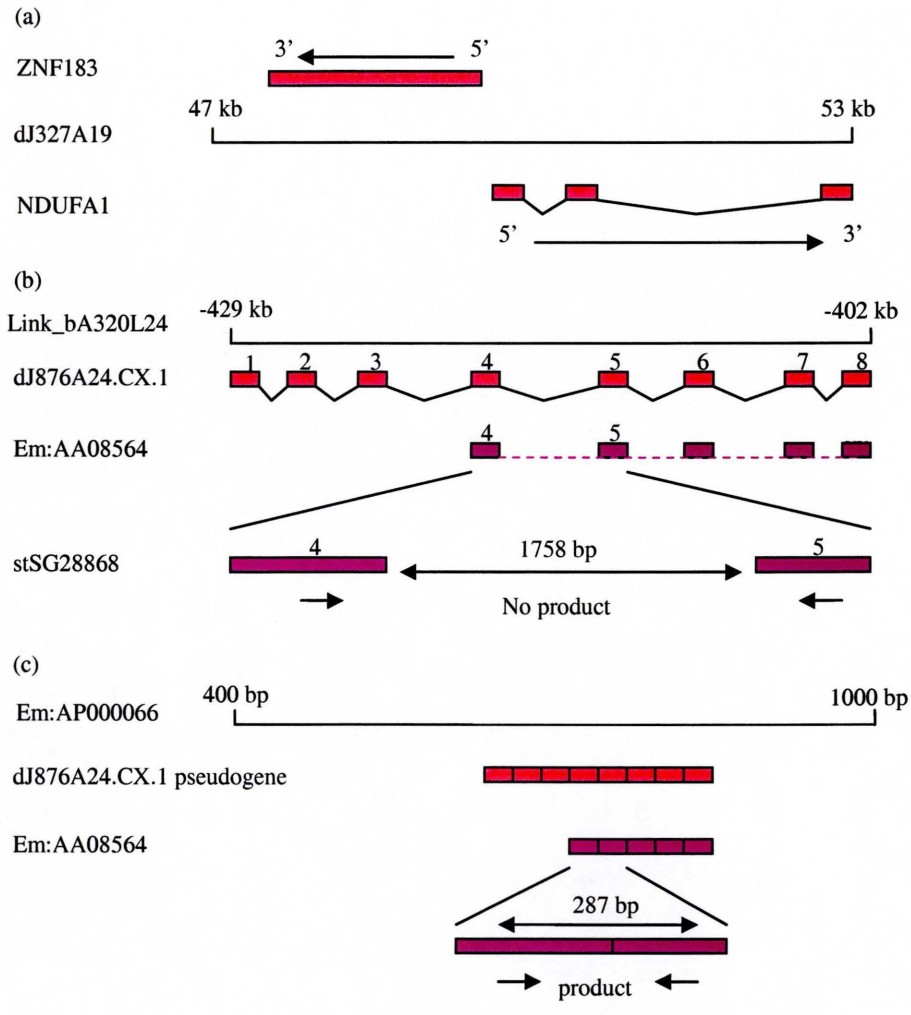
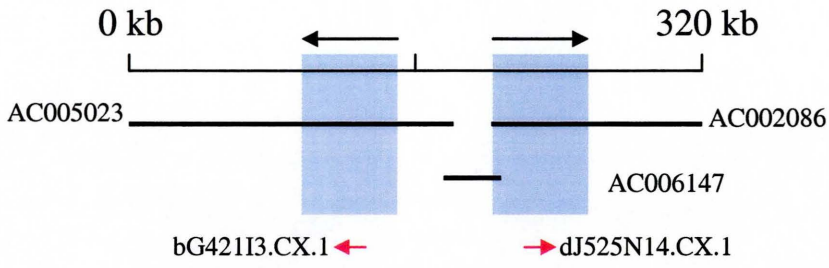


Figure 4.12: *Genes in their genomic context (2) (a) ZNF183 (shown as a single red box above the solid line) is transcribed in the opposite orientation to NDUFA1 (shown as red boxes linked by black lines below the line). The position of the regulatory elements for each gene is unknown. (b) dJ876A24.CX.1 (shown as red boxes linked by black lines) was identified in part with an EST (Em:AA085642 shown as purple boxes). The expansion of the alignment of the EST to exons four and five and the position of the primers for stSG28868 is shown. Intron five is 1758 bp and RH mapping using stSG28868 did not reveal a location on the X chromosome. (c) A pseudogene of dJ876A24.CX.1 (shown as red bars) on Chromosome 8 contains no introns and the alignment of the EST (shown as purple bars) generates a product of 287 bp, and the RH mapping did reveal a location on Chromosome 8.*

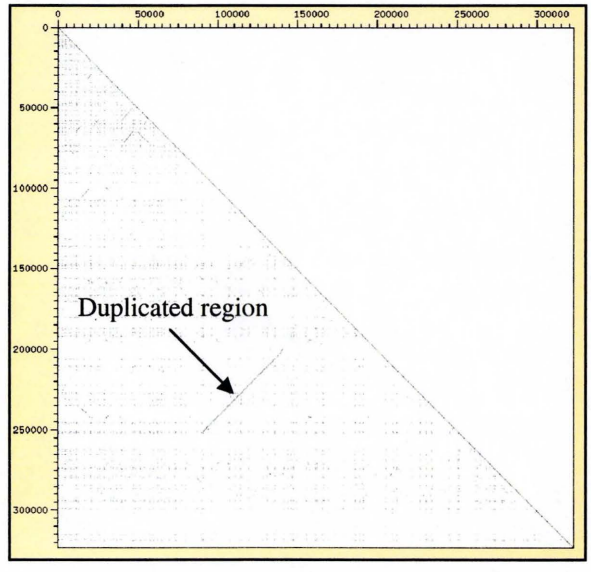
As was discussed in the previous chapter (Section 3.4), the availability of the genomic sequence allows for the identification of previously uncharacterised low copy repeats. There is an inverted repeat of approximately 50 kb within the region and two very similar genes, dJ525N14.CX.1 and bG421I3.CX.1, are present within the repeat (see Figure 4.13). The predicted amino acid sequence of both genes shows that there are only two amino acid differences between them (Asn93 to Asp and Cys151 to Arg). The cDNA sequence generated to confirm the gene structures matched exactly to dJ525N14.CX.1. The lack of supporting evidence for bG421I3.CX.1 and the non-conservative nature of the amino acid differences does not mean that the gene is not functional. It may be expressed at much lower levels or in a limited number of tissues, and confirmation may be obtained by screening a wider variety of cDNA resources.

Figure 4.13: *(see over) Analysis of 50 kb duplication (a) Inverted duplication of a 50 kb region (shown as blue boxes, black arrows indicate orientation). The sequence contribution of each clone is shown as black bars, and EMBL accession numbers of the sequences that make up the region are given. The position of the each gene and the direction of transcription are shown as red arrows. (b) A DOTTER (Sonnhammer, E. L., et al., 1995) of the region matched against itself, the continuous black line along the diagonal is the match against itself, and the smaller black line perpendicular to the central diagonal indicates the position of the duplication. (c) The alignment of the predicted protein sequences of the two genes, dJ525N14.CX.1 and bG421I3.CX.1. Identifical matches are shown as black boxes, the two amino acids that differ (* N to D and ^ C to R) are shown as white boxes.*

(a)



(b)



(c)

```
bG421I3.CX.1      1 MEPPDQCSQYMTSLLSPAVDDEKELQDMNAMVLSLTEEVKEEEEDAQPEPEQGTAAGEKL
dJ525N14.CX.1     1 MEPPDQCSQYMTSLLSPAVDDEKELQDMNAMVLSLTEEVKEEEEDAQPEPEQGTAAGEKL

                                                  *
bG421I3.CX.1     61 KSAGAQGGEEKDGGGEEKDGGGAGVPGHLWEGNLEGTSGSDGNVEDSDQSEKEPGQQYSR
dJ525N14.CX.1    61 KSAGAQGGEEKDGGGEEKDGGGAGVPGHLWEGDLEGTSGSDGNVEDSDQSEKEPGQQYSR

                                                 ^
bG421I3.CX.1    121 PQGAVGGLEPGNAQQPNVEAFTPLQLQELECIFQREQFPSEFLRRRLARSMNVTELAVQI
dJ525N14.CX.1   121 PQGAVGGLEPGNAQQPNVEAFTPLQLQELERIFQREQFPSEFLRRRLARSMNVTELAVQI

bG421I3.CX.1    181 WFENRRAKWRRHQRALMARNMLPFMAVGQPVMVTAAEAITAPLFISGMRDDYFWDHSHSS
dJ525N14.CX.1   181 WFENRRAKWRRHQRALMARNMLPFMAVGQPVMVTAAEAITAPLFISGMRDDYFWDHSHSS

bG421I3.CX.1    241 SLCFPMPPFPPPSLPLPLMLLPPMPPAGQAEFGPFPFVIVPSFTFPNV
dJ525N14.CX.1   241 SLCFPMPPFPPPSLPLPLMLLPPMPPAGQAEFGPFPFVIVPSFTFPNV
```

* N (Asparagine) to D (Aspartate) ^ C (Cysteine) to R (Arginine)

4.4. Predicting the function of novel gene products

One of the major challenges once genes have been identified is to predict the role they play within the cell. In order to ascertain the function of the genes experimentally, some initial hypothesis of potential function would greatly facilitate the experimental investigation. As discussed in Section 1.5, there are a variety of methods available to gain an insight into the particular function of novel genes. As part of the systematic identification of the genes, genomic sequence is aligned to previously generated nucleotide and protein sequence. Genes that show similarity to genes of previously determined function may have a related function. In some cases, novel genes are not similar to any genes for which function has been previously determined, and for these, *ab initio* prediction of function can be carried out. For instance, proteins are made up of functional units or domains, and identification of homologues of well-known functional domains with novel proteins may predict specific biochemical functions which may be ascribed to the novel protein.

The 44 genes identified in the 8 Mb region between DXS7598 and DXS7333 have been functionally characterised as far as the available information allows (see Table 4.4). The function of eleven genes within the region has been previously deduced experimentally by others (see reference column in Table 4.4). For instance, the Glutamate Receptor Subunit 3 gene (GRIA3) is a member of the glutamate receptor protein family that mediates most of the excitatory neurotransmission in the mammalian brain (Gecz, J., *et al.*, 1999). The function of a further ten genes is inferred based on their similarity at the protein level to genes whose function has been determined experimentally. For example, Septin2 is a member of the septin family of

genes and is an orthologue of the KIAA00128 gene in rat. It has been shown that this rat homologue of the Septin2 is one of four septin proteins that form a filament around which actin bundling can occur (Kinoshita, M., *et al.*, 1997). The function of two of the fourteen novel genes identified and confirmed in this chapter (see Section 4.2) has been inferred based on their similarity to genes of known function. The first novel gene, dJ327A19.CX.4 is 70 % identical at the protein level to a rat gene, testis-specific A-kinase anchoring protein (TAKAP-80), which is involved in sperm motility by binding to a type II cAMP-dependent protein kinase which tightly associated with the fibrous sheath (Mei, X., *et al.*, 1997). The second novel gene with inferred function is dJ318C15.CX.1, which is 90% identical at the protein level to the human gene, cullin 4A (Osaka, F., *et al.*, 1998). Cullins are hydrophilic proteins found in yeast, worm and human that are involved in cell cycle regulation through protein degradation (Mathias, N., *et al.*, 1996).

Table 4.4: *Functional characterisation of Genes*

Gene Name	Characterisation	Function or function prediction	Reference
HOM-TES-85	Known	HOM-TES-85 tumour antigen	direct submission
hATB0+	Known	Amino acid transporter	Sloan, J. L., *et al.*, 1999
ANT2	Known	Adenine nuceotide translocator	Giraud, S., *et al.*, 1998
NDUFA1	Known	Accessory protein in mitochondria complex I	Au, H. C., *et al.*, 1999
LAMP2	Known	Lysosomal associated membrane protein	Kannan, K., *et al.*, 1996
GLUD2	Known	Glutamate dehydrogenase	Shashidharan, P., *et al.*, 1994
GRIA3	Known	Neurotransmitter receptor	Gecz, J., *et al.*, 1999
T-plastin	Known	Actin bundling protein	Lin, C. S., *et al.*, 1999
NRF	Known	Transcription regulation	Nourbakhsh, M., *et al.*, 2000
IL13R	Known	Cell surface receptor, binds IL13	Aman, M. J., *et al.*, 1996
ZNF-kaiso	Known	Zinc finger transcription factor	Daniel, J. M., *et al.*, 2001
HPR6.6	Inferred	Steroid binding membrane protein	Gerdes, D., *et al.*, 1998
ZNF183	Inferred	C3HC4 Ring finger	Frattini, A., *et al.*, 1997
UBE2A	Inferred	Orthologue of yeast RAD6, a DNA repair enzyme	Koken, M. H., *et al.*, 1996
ATP1B4	Inferred	K-ATPase beta sub-unit	Pestov, N. B., *et al.*, 1999
SMT3B	Inferred	Kinetochore associated protein	Lapenta, V., *et al.*, 1997
Septin2	Inferred	Septin filament protein	Kinoshita, M., *et al.*, 1997
RPL39	Inferred	Ribosomal protein	Delbruck, S., *et al.*, 1997
U69a	Inferred	Small nucleolar RNA	direct submission
dJ327A19.CX.4	Inferred	similar to rat testis-specific A-kinase anchoring protein	Mei, X., *et al.*, 1997
dJ318C15.CX.1	Inferred	Homologue of cullin-4A	Osaka, F., *et al.*, 1998
bA161I19.CX.1	Interpro prediction	Leucine rich repeat (IPR001611)	-
bA161I19.CX.2	Interpro prediction	EF-hand (IPR002048), 2 Calporin homology (IPR001715)	-
dJ169K13.CX.1	Interpro prediction	6 kelch repeats (IPR001798), 1 BTB/POZ (IPR000210	-
dA155F9.CX.1	Interpro prediction	Plekstrin homology (IPR001849)	-
dJ562J12.CX.2	Interpro prediction	Zinc finger CCHC type (IPR001878)	-
dJ555N2.CX.1	Interpro prediction	6 bipartite nuclear localisation signals (IPR001472)	-
dJ1139I1.CX.1	Interpro prediction	3 mitochondrial energy transfer (IPR001993	-

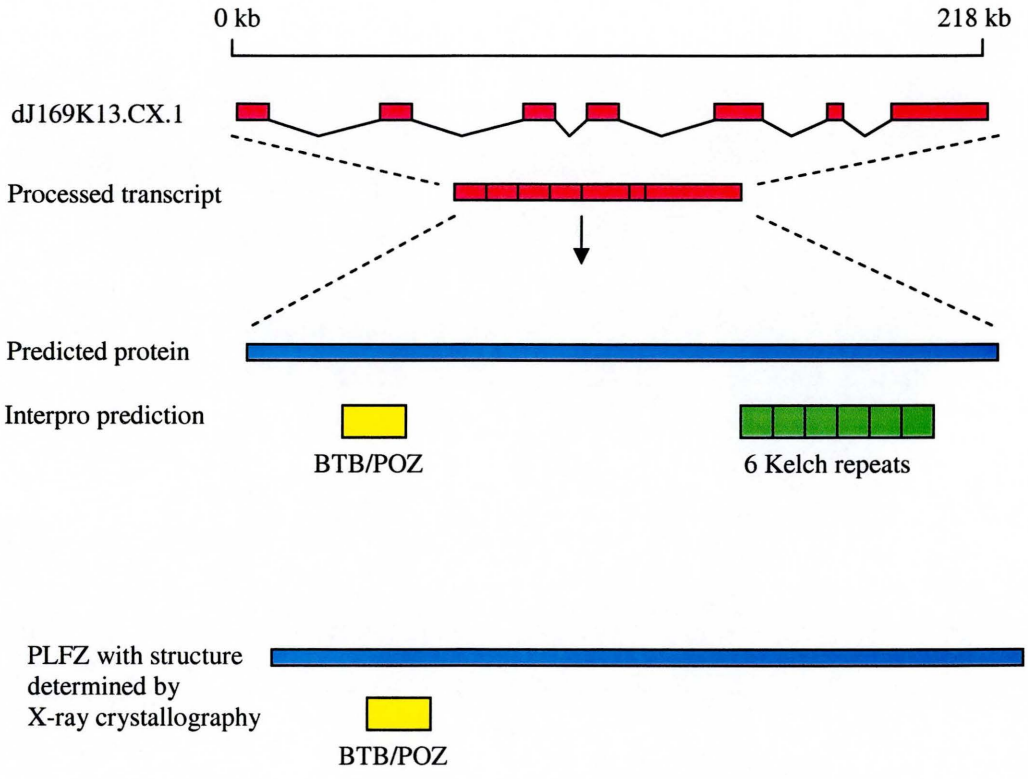
UPF3b	Interpro prediction	5 bipartite nuclear localisation signals (IPR001472)	-
bG421I3.CX.1	Interpro prediction	Homeobox (IPR001356), proline rich (IPR000694)	-
bG421I3.CX.3	Interpro prediction	Homeobox (IPR001356)	
dJ525N14.CX.1	Interpro prediction	Homeobox (IPR001356), proline rich (IPR000694)	-
dJ876A24.CX.1	Tmpred	Transmembrane protein	-
bB115H14.CX.1	Unknown	-	-
bB125M24.1	Unknown	-	-
dJ170D19.CX.2	Unknown	-	-
dJ327A19.CX.3	Unknown	-	-
bG421I3.CX.2	Unknown	-	-
dJ525N14.CX.3	Unknown	-	-
bA45J1.CX.1	Unknown	-	-
dJ655L22.CX.1	Unknown	-	-
MCT-1	Unknown	-	-
dJ321E8.CX.2	Unknown	-	-
dJ321E8.CX.3	Unknown	-	-

The remaining 23 genes do not appear to show any similarity to genes of known function, and these have been analysed for the presence of protein domains. This analysis was carried out using INTERPRO, a web-based tool that uses a series of protein domain prediction programmes (see http://www.ebi.ac.uk/interpro/scan.html) which compare novel protein sequence to sequences of known protein domains.

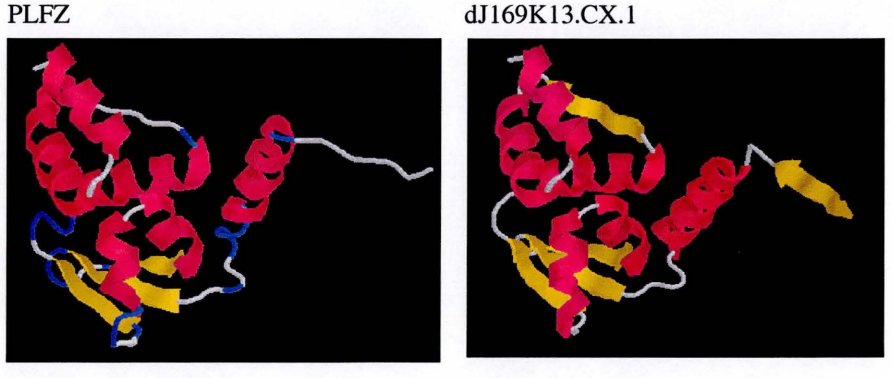
The analysis shows that eleven genes are predicted to contain one or more previously defined protein domains and an example of this is shown in Figure 4.14. One gene, dJ169K13.CX.1 is predicted to contain a BTB/POZ domain and 6 kelch repeats. The kelch repeat was first identified in the kelch gene from *Drosophila*, and repeating kelch units form anti-parallel beta-sheets that come together to form a propeller like structure. Proteins containing six-kelch repeats have been shown to have a role in actin bundle formation (e.g. scruin, Way, M., *et al.*, 1995), and this may suggest a possible role for dJ169K13.CX.1.

Figure 4.14: *(see over) Functional analysis of genes (a) Identifying a possible function for dJ169K13.CX.1 (shown as red boxes linked by black lines). The processed transcript (shown as red boxes) is translated to generate the predicted protein sequence (shown as a blue box), which is compared to known protein and their domains. dJ169K13.CX.1 is predicted to contain a BTB/POZ domain (shown as a yellow box) and 6 kelch repeats (shown as green boxes). PLFZ, a protein whose structure has been determined by X-ray crystallography also contains a BTB/POZ domain and the two sequences corresponding the BTB/POZ domain were aligned and the structure of the domain within dJ169K13.CX.1 was predicted using MODELLER. (b) The predicted structure of the BTB/POZ domain of dJ169K13.CX.1 compared to the known structure of PLFZ (viewed in RASMOL). The domain is made up of a series of alpha helices (shown in red) flanked by beta sheets (shown in yellow). Looping regions are shown as either white or blue lines.*

(a)


0 kb
218 kb
dJ169K13.CX.1
Processed transcript
Predicted protein
Interpro prediction
BTB/POZ
6 Kelch repeats
PLFZ with structure
determined by
X-ray crystallography
BTB/POZ

(b)


PLFZ
dJ169K13.CX.1

The BTB/POZ domain (BTB = BR-C, ttk and bab (Zollman, S., *et al.*, 1994), and POZ = Pox virus and Zinc finger (Bardwell, V. J., *et al.*, 1994) is a much simpler domain, comprising of a cluster of alpha-helices flanked by short beta-sheets, determined using X-ray chrystallography on the promyelocytic leukemia zinc finger oncoprotein (plzf) (Ahmad, K. F., *et al.*, 1998). The BTB/POZ domain mediates homomeric and in some cases heteromeric dimerisation. An alignment of the BTB/POZ domains from plfz and dJ169K13.CX.1 was carried out using CLUSTALW (data not shown). The structure of the BTB/POZ domain in dJ169K13.CX.1 was then predicted using MODELLER which uses experimentally determined protein sequences to predict conformation of the other proteins with similar amino acid sequences (Sanchez, R., *et al.*, 1997). This is possible because a small change in the sequence usually results in a small change in the 3D structure. The predicted structure for the BTB/POZ domain of dJ169K13.CX.1 is shown in Figure 4.14b.

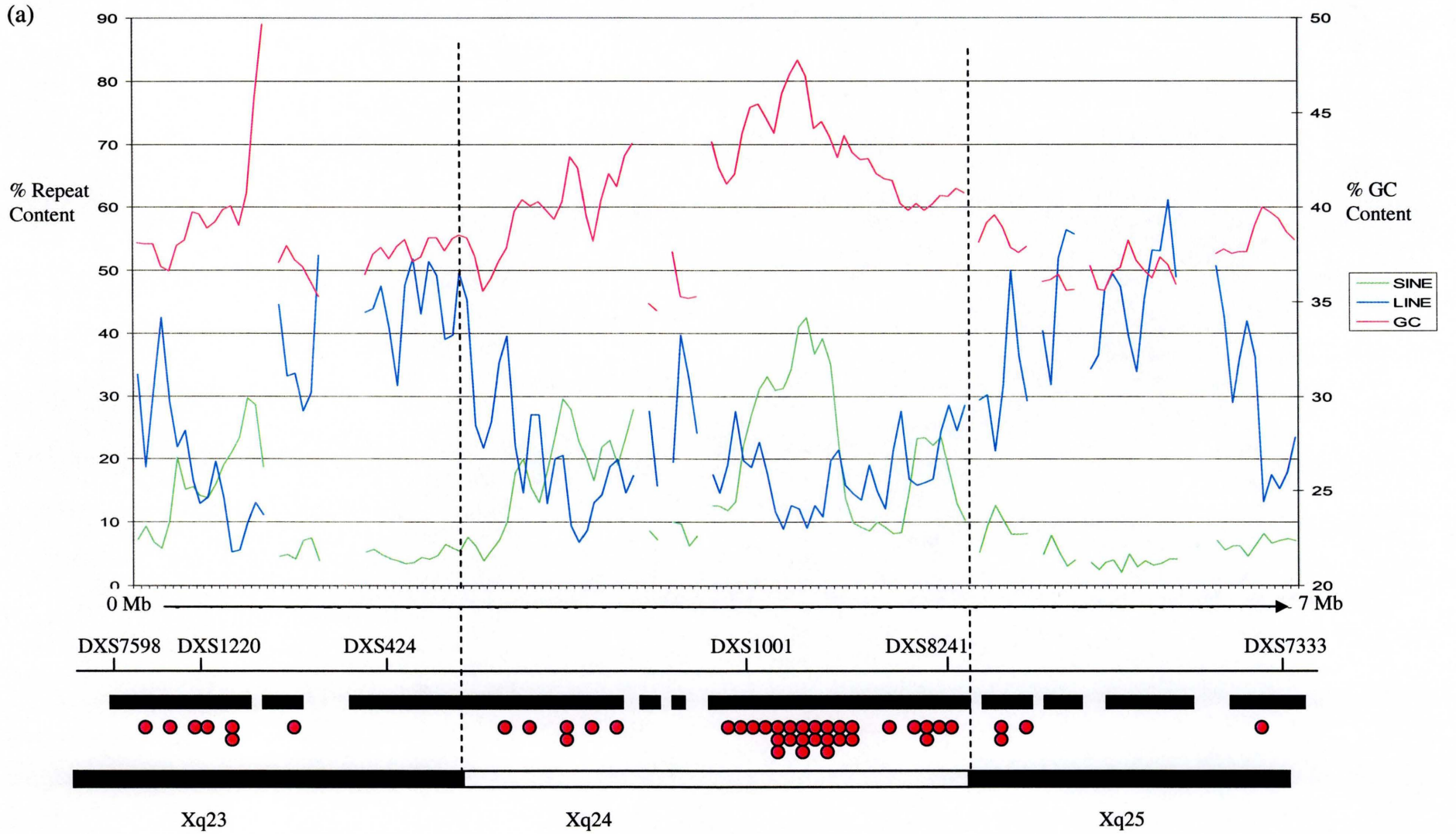
4.5 Analysis of the sequence composition of the region in Xq23-Xq24

The availability of the genomic sequence and the identification of the genes contained within allow for a detailed analysis of the correlation between the overall sequence composition of the region, the distribution of the repetitive elements and the gene density. It also allows for a comparison with other regions of the human genome. Previous mapping and cytogenetic analysis predicted the region of interest contains the distal portion of Xq23, part of a G light band, and the whole of Xq24, a complete R dark band. As discussed in section 3.4, R-bands are thought to be generally gene rich and SINE rich, while G-bands are thought be generally gene poor and SINE poor. The availability of the genomic sequence and the identification of genes contained within allows for a detailed analysis to be carried out, to determine the relationship between gene content and sequence composition.

The twelve sequence contigs in the region were analysed by dividing each sequence contig into 100 kb segments that overlapped by 50 kb. Each segment was then analysed for the repeat content and GC content using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) and the results plotted as single points on a graph (see Figure 4.15a). The positions of the genes relative to the sequence segments were identified. The lines indicate marked alterations in GC content, LINE and SINE content. These would be consistent with possible location of boundaries defined cytogenetically assuming one or more of these sequence features affects intensity of Giemsa staining. The GC content is highest in the region predicted to be the light band Xq24, and coincides with the highest density of genes. The areas

of high SINE content and low LINE content correspond with a relatively high gene density. This observed correlation between SINE content and gene density agrees with observations carried out over the whole genome (IHGSC, 2001).

Figure 4.15: *(a) (see over) Genome landscape of the region of interest. The X-axis represents the extent of the region divided into 100 kb segments, overlapping by 50 kb. The Y-axis on the left is the % repeat content, and the one on the right is the % GC content. Gaps in the plot indicate gaps in the finished sequence. The lines indicate marked alterations in GC content (red), LINE (blue) and SINE content (green). The genomic interval is represented as single thin black line below the chart, with key markers positoned. The extent of sequencing is shown as black boxes and the position of each gene is represented by a red circle. The hypothetical positions of the cytogenetic bands are also given (black dotted lines). In general, there is good correlation between SINE content and gene density.*


(a)
% Repeat Content
% GC Content
SINE
LINE
GC
0 Mb
7 Mb
DXS7598
DXS1220
DXS424
DXS1001
DXS8241
DXS7333
Xq23
Xq24
Xq25

Comparisons were also carried out with two other light bands on the X chromosome (Xp11.23 and Xq26.1) as well as with the whole of the X chromosome, chromosome 22 and the whole genome (see Figure 4.15b). For each X chromosome band, a plot of the GC and repeat content was generated as was described for Xq23-Xq24 and average figures for GC content, SINE, LINE and particularly L1, given the proposed role LINE elements are thought to play in X inactivation (Lyon, M. F., 1998 and see Section 1.5). The extent of each band was identified using sequence content (high GC and SINE content) and the expected position of the band based on previous mapping information (FISH data and gene and marker placement). The average figures for the whole of the X chromosome and the whole genome were obtained from analysis of sequence in ENSEMBL (courtesy of Ewan Birney and Mark Ross) and figures for chromosome 22 were obtained from published sources (Dunham, I., *et al*, 1999).

Xq24 has an average GC content of 40% and gene density of 10 per megabase (taking 33 genes in the 3.3 Mb). In comparison, chromosome 22 has a much higher GC content (48%) and a higher gene density (22 genes per megabase). Xq24 has a relatively high SINE content of 21% when compared to the rest of the genome (13%), as expected, as the figures for the whole genome includes regions within R-band which are thought to be gene poor and SINE poor. However, Xq24 has a much lower SINE content when compared to Xp11.23 (33%), which is considered to be one of the most gene rich regions on the X chromosome (IHGSC, 2001). Although chromosome 22 is considered to be a gene-rich chromosome, and SINE content is thought to be linked to gene density, the average SINE content on chromosome 22 (17%) is less than that of both Xq24 (21%) and Xp11.23 (33%). However, chromosome 22 is gene rich when compared to other whole chromosomes, not individual light bands, and

therefore as expected has a higher SINE content when compared to the average figures for the whole of the X chromosome (11%).

The Lyon hypothesis suggests an involvement of L1 elements in X-inactivation and hypothesises a higher LINE1 content on the X chromosome than for the other regions of the human genome (as discussed in Section 1.5). It has been shown that the X chromosome in general has a much higher L1 content (30%) when compared to the average figure for the whole genome (17%) (IHGSC, 2001). Xq24 has a higher L1 content (13%) than that reported previously for Chromosome 22 (9%) supporting Lyon hypothesis that there will be more LINE1 elements on the X chromosome.

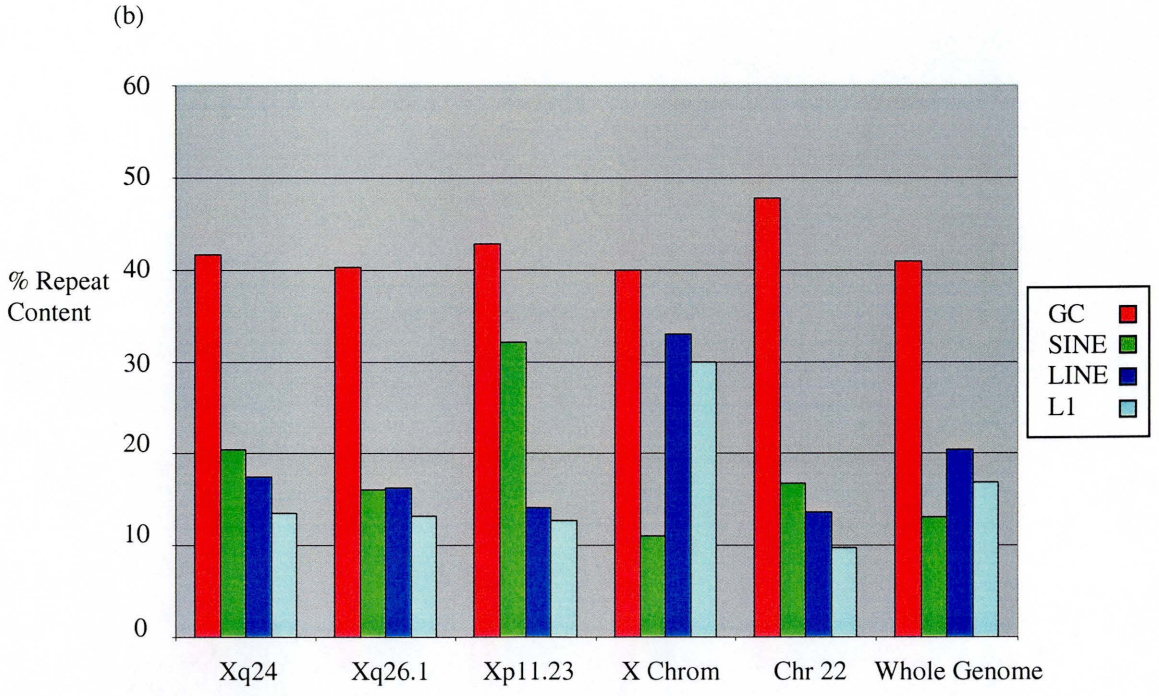


Figure 4.15: *cont. (b) Comparisons of genome landscapes of other regions of the genome. The average GC content (red bars), SINE content (green bars), LINE content (dark blue bars), and LINE1 content (light blue bars) across the region has been calculated. Assuming there is a correlation between SINE content and gene density, Xq24 appears slightly more gene dense than Xq26.1, but much less gene dense than Xp11.23, one of the gene dense regions of the X chromosome (IHGSC, 2001) . There appears to be a higher LINE1 content in the X chromosome bands than on Chromosome 22, which agrees with the hypothesis that LINE1 elements maybe involved in X-inactivation.*

4.6 Mutation screening for MRX23

The identification of genes within a given region provides a valuable resource in the search for disease causing mutations. The critical regions for a number of diseases for which no causative mutation has been identified, are contained within or overlap with the 8 Mb region between DXS7598 and DXS7333 (see Figure 4.16). The nature of X-linked non-specific mental retardation (MRX) syndromes (the absence of any distinguishing phenotype other than mental retardation) results in the assignment of a different MRX number to each family diagnosed with the condition. MRX23 (discussed further in Section 1.4.3) has previously been localised to a region of the X chromosome between DXS1220 and DXS424, a region of approximately 2.4 cM (Gregg, R. G., *et al.*, 1996). Analysis of the available sequence shows that these markers have now been accurately positioned. DXS1220 is located within Chr_Xctg20, 500 kb from the distal end, and DXS424 is located within Chr_Xctg3, 150 kb from the proximal end. The gap between the two contigs has been sized by fibre fish to approximately 300-500 kb (see Figure 4.9).

Work is currently underway to generate sequence across the gap between the two contigs as part of the X chromosome mapping and sequencing project. As a result the critical region for MRX23 can now be more precisely defined, with boundaries accurately placed on the genomic sequence. Allowing for upper and lower estimates of the size of the remaining gap, the critical region is thought to be between 950 kb and 1150 kb. To date, the critical region between DXS1220 and DXS424 has been shown to contain 3 genes, $hATB^{0+}$, T-plastin and SMT3B. $hATB^{0+}$ has been shown to be an amino acid transporter, with the particular affinity for hydrophobic amino acids

(Sloan, J. L., *et al.*, 1999). T-plastin is a member of the plastin family of actin binding proteins and is expressed in most tissues in higher eukaryotes (Arpin, M., *et al.*, 1994, Lin, C. S., *et al.*, 1993). Although little is known about the function of SMT3B, it shares 87% amino acid identity with SMT3A, and both are thought to be human homologues of the yeast SMT3, which are essential in chromosome segregation (Lapenta, V., *et al.*, 1997). Genes with a wide variety of function are associated with mental retardation-related disorders therefore all genes were considered to be positional candidates and screened for disease-causing mutations. Mutation screening of all three candidate genes was carried out by amplification of the exons using DNA from an affected individual and DNA from an unaffected, unrelated individual and sequencing the products.

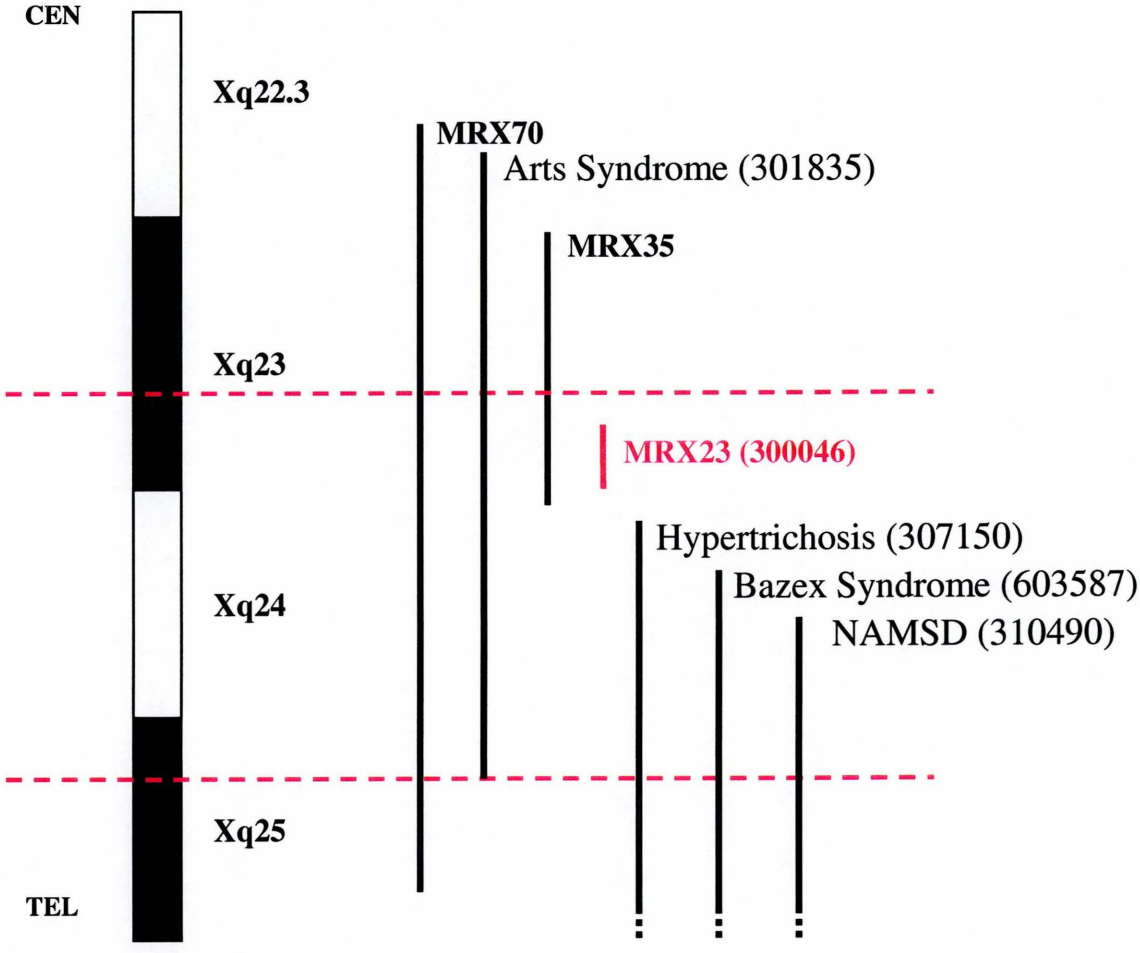


Figure 4.16: *Uncloned diseases mapping to region of interest. A section of the X chromosome showing the extents (vertical black lines) of the critical regions of uncloned diseases that include the region of interest between DXS7598 and DXS7333 (between the two dotted red lines). The critical region for MRX23 is shown in red. OMIM numbers, where available are given in brackets. For information on MRX families see Toniolo, D., et al., 2000.*

A total of 31 primer pairs were designed to amplify each of the exons of the three genes including at least 50 bp of flanking intron sequence (see Table 4.7 in appendix to this chapter). The PCR was used to amplify DNA from one affected individual (kindly provided by Ron Gregg, see Section 2.7) and one unaffected, related individual. The products were excised, purified and sequenced (for example see Figure 4.17). In each experiment, a control STS lying close to the candidate gene was amplified as a positive control in the case of a deletion of an exon or exons where no product would be observed. To date, sequences from all 31 exons have been compared and only one difference has been observed between affected and unaffected individuals.

A single base difference, T to A, at nucleotide position 36 of exon 2 of the $hATB^{0+}$ gene was identified in the affected individual. This is a synonymous change as it alters GGT to GGA, both of which code for glycine (see Figure 4.18). This alteration in the sequence does not appear to cause the formation of a cryptic splice site, the single change does not appear to introduce a consensus sequence for a novel 5' or 3' splice site. Analysis of other sequence aligned to the genomic sequence in the region (cDNA and EST sequence) shows that only DNA sequence generated from the affected individual contains GGA, and all other sequence contained GGT.

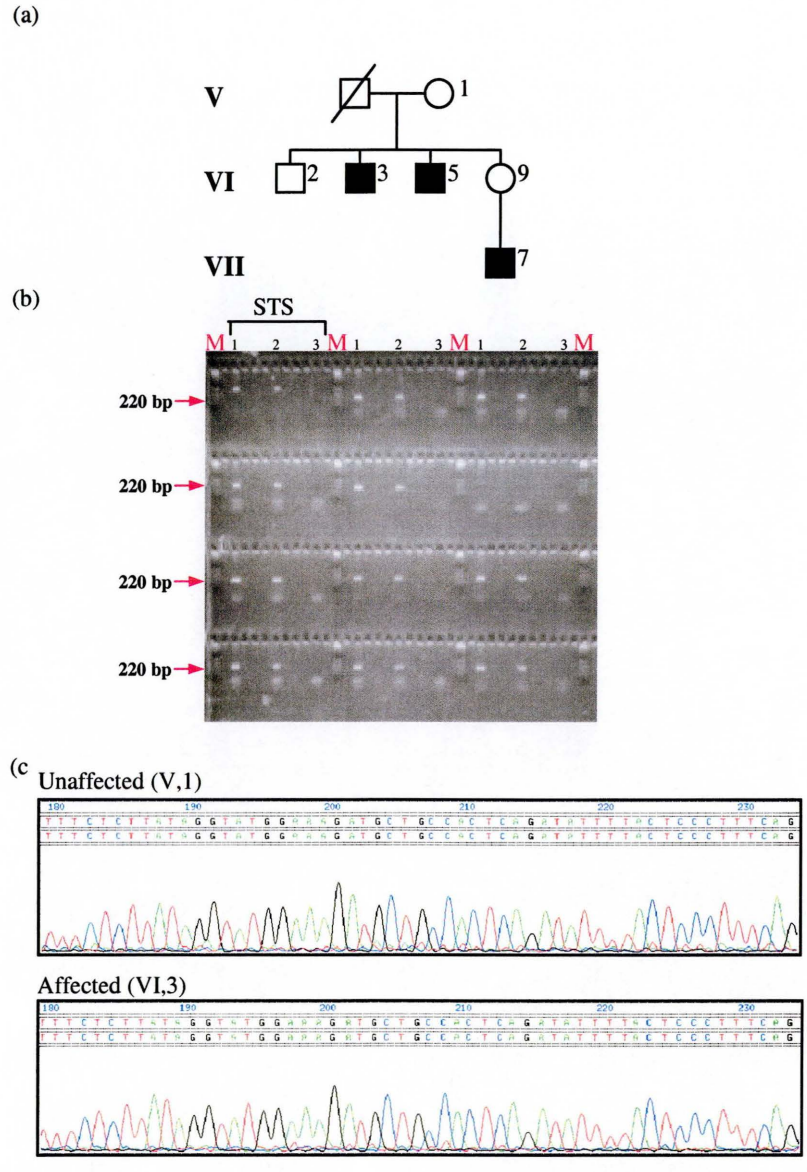


Figure 4.17: *Mutation screening for MRX23 (a) Part of the pedigree for the MRX23 family indicating the relationships between unaffected and affected family members (Males = squares, females = circles, affected = shaded black, deceased = line through). DNA from VI,3 was used for mutation screening. (b) Twelve STSs designed to amplify exons 1-12 of the T-plastin gene, screened against DNA from affected and unaffected individuals (lane 1 = unaffected, lane 2 = affected, lane 3 = $T_{0.1}E$, M = marker). (c) The sequence of part of exon 4 from both an affected and an unaffected family member. Alignment of the two sequences reveals no differences.*

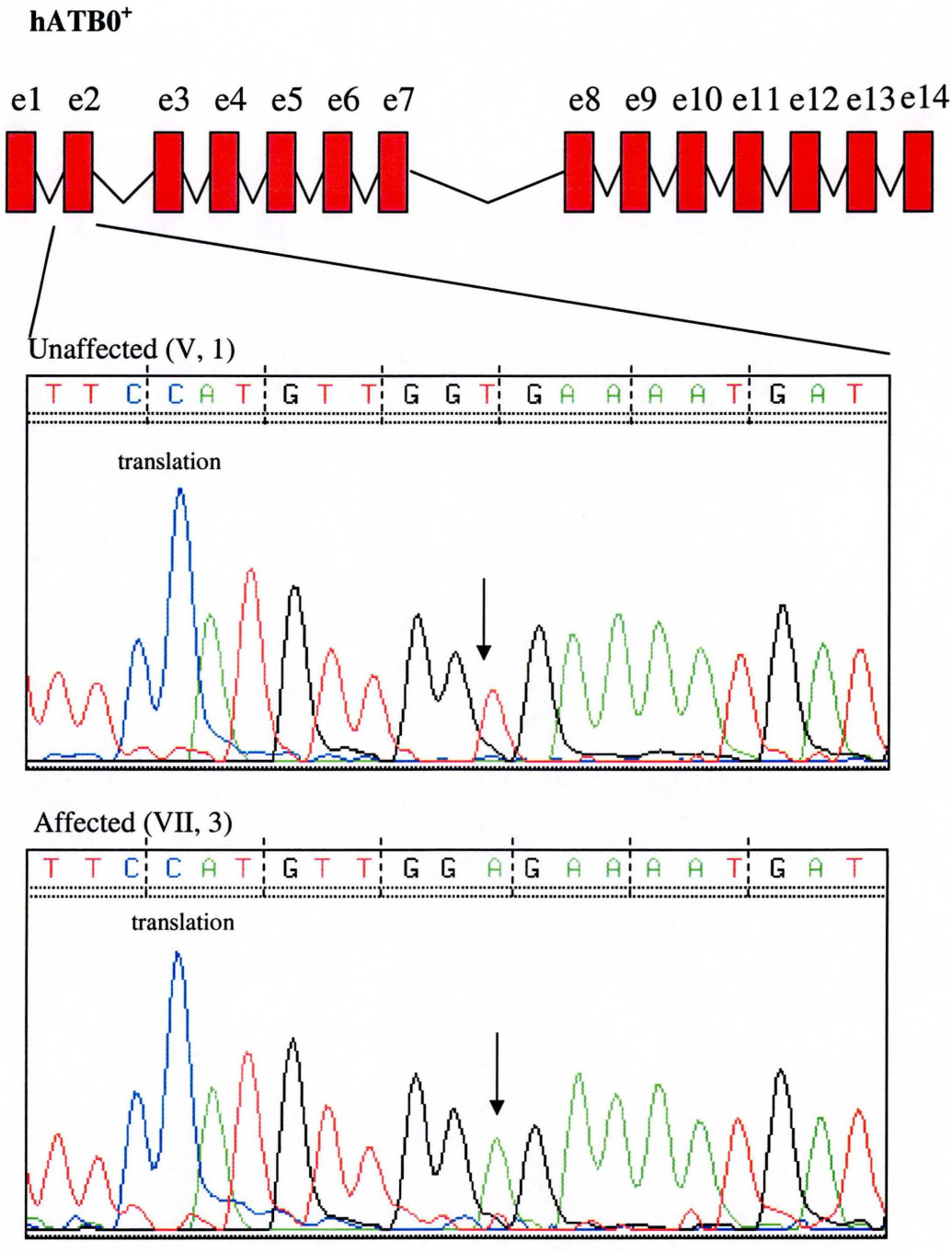


Figure 4.18: *Identification of a potential silent mutation. Top of the figure shows a schematic of the the hATB0+ gene (exons shown as red boxes, introns shown as 'v'-shaped lines). The bottom of the figure shows a potential silent mutation (GGT to GGA, indicated by the black arrow on the sequence (viewed in TREV) from both the unaffected and the affected) was observed within exon two of the hATB0+ gene (shown as red boxes linked together with black lines).*

A more detailed analysis has not been carried out while the sequence of the region remains incomplete and therefore while one or more other candidate genes may be discovered and included in the first phase of mutation detection. Further analysis of the GGT to GGA variant would involve testing other members of the family, as well as other unaffected individuals to exclude the possibility it is a polymorphism in the population and that the change is specific to the MRX23 family. If the polymorphism did appear to segregate with the affected individuals a functional assay could be carried out to to identify any functional implications of the variant.

4.7 Discussion

A gene map covering approximately 8 Mb of Xq23-24 between DXS7598 and DXS7333 has been constructed and contains 33 confirmed genes of which 14 have been described in this chapter, 11 genes predicted by a series of similarity searches and gene prediction packages and 20 pseudogenes. Contiguation of the bacterial clone contigs covering the region and subsequent completion of the genomic sequence, will provide the basis for the identification of further novel genes in this region. The largest sized gap in the region lies within the critical region for MRX23 and is thought to be approximately 500 kb.

Where possible, the entire predicted ORF for each gene has been confirmed in cDNA sequence. The methods described here for gene identification require the presence of cDNA sequence to confirm a predicted gene. A number of different cDNA sequencing projects around the world (for a full list see Table 4.8 in appendix to this chapter) are depositing cDNA sequence into public databases and Figure 4.19 shows

the relative contribution of each collection to the confirmation of genes in the region studied. This shows that random cDNA sequencing can greatly facilitate the identification of genes.

The gene structures predicted by GENSCAN and FGENESH have been compared to the genes that were confirmed by cDNA or EST sequence using BLAST (Altschul, S. F., *et al.*, 1990), and GENSCAN correctly predicted 42.4 kb of the 80.9 kb of expressed bases (52%), whereas FGENESH correctly predicted 40.5 kb (50%) (see Figure 4.20a – red bars). As gene prediction programs only predict coding sequences, the figure is artificially lower due to the inclusion of the untranslated regions in the total amount of expressed sequence. Figure 4.19b shows two examples of genes predicted using GENSCAN and FGENESH. In general, they are better at predicting internal exons but are unable to predict the 3' and 5' most exons. This is consistent with analysis carried out on larger gene sets (e.g Chr22 gene set, Dave Beare personal communications, The Sanger Institute).

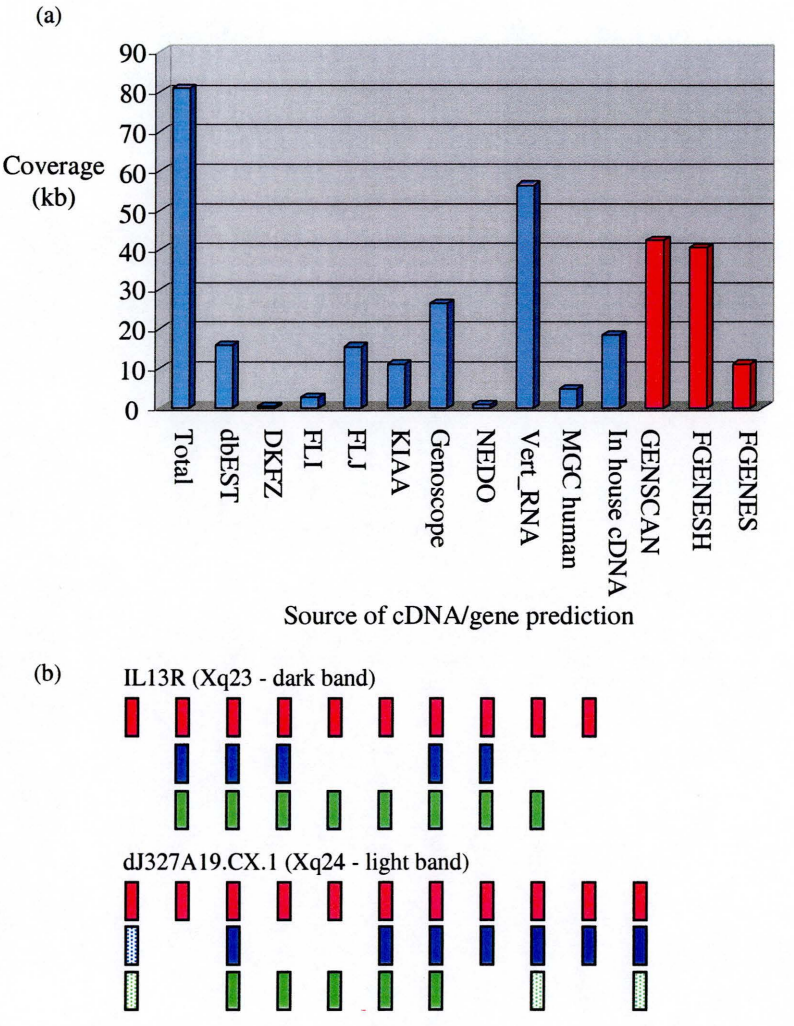


Figure 4.19: *Contributions of cDNA sequencing projects and prediction programs (a) Breakdown of the relative contribution of each cDNA resource (blue bars) and each gene prediction package (red bars). The total amount of expressed sequence (first bar) is approximately 80 kb. The vertebrate mRNA database (vert_mRNA) provides over 50 kb and 18 kb of cDNA sequence was generated in house. GENSCAN and FGENESH predict around 50% of the expressed sequence. This is slightly lower than expected as the total figure for expressed sequence includes UTR, regions the gene prediction programs do not predict. (b) Two examples of genes (exons shown as red boxes) predicted by GENSCAN (blue boxes) and FGENESH (green boxes). Partially predicted exons are represented by hashed boxes.*

The region of interest was contained in twelve sequence segments covering 7 Mb. A further 600 kb of sequence was available as unfinished sequence. Analysis of the unfinished sequence revealed no new genes and one pseudogene. This finding is not suprising as 90% of the draft sequence was predicted to be within gene-poor regions at the proximal portion of Xq25. In two cases though, the draft sequence contained exons from genes placed on the finished sequence (GRIA3 and Serotonin-5HTR2), and these genes both cover large regions of genomic sequence. Therefore one of the disadvantages of draft sequence is that the larger gene structures, the genes with large introns are more likely to be present on multiple draft sequence contigs.

As discussed in Section 4.3, it is difficult to identify the complete ORF and flanking UTR sequences for each gene. In this study a total of 22 primer pairs, designed to regions thought to contain genes, failed to identify any positive pools in the cDNA libraries. In one case, a PCR assay was designed to generate cDNA sequence at the 5' end of a gene (see Figure 4.20a). In a second example, the primer pair was designed to an exon of a gene predicted by both GENSCAN and FGENESH, and with a BLASTX homology (see Figure 4.20b). Both PCR assays failed to identify any positive pools in the libraries available. The absence of such confirmation for predicted genes does not mean the gene is not real. It may be expressed in a tissue that is not represented in the cDNA collection tested or at low levels, or for a short period of time, for instance during development. Screening of cDNA libraries from a wider variety of human tissues will increase the likelihood of confirming a particular gene but one complementary strategy (discussed in chapters five and six) is to use comparative sequence analysis.

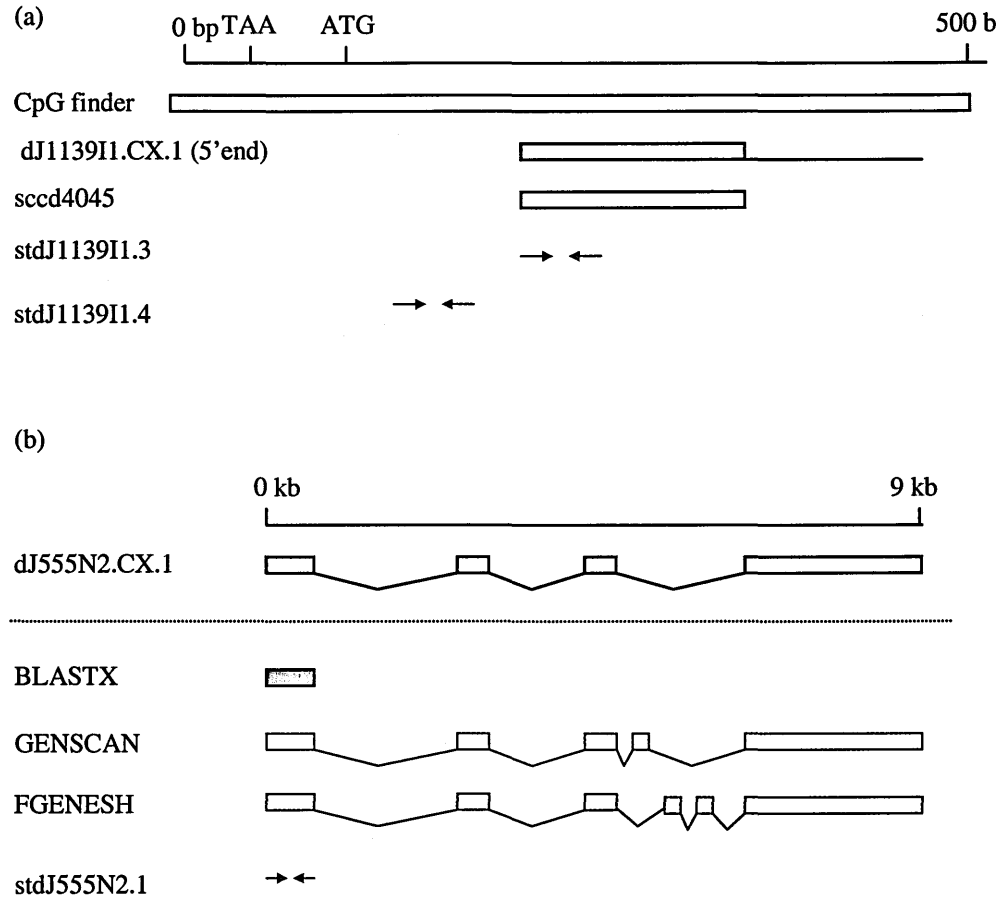


Figure 4.20: *Examples of unconfirmed genes (a) The 5' end of dJ1139I1.CX.1 (shown as a red box), the extent of the cDNA sequence confirming the 5' end of the gene is shown as pink boxes. The position of the most likely translation start site (ATG) is shown, a well as the first in-frame stop codon. Primers were designed (shown as red arrows), both within the cDNA sequence and just upstream but no 5' extension of the sequence was observed. (b) dJ555N2.CX.1 (shown as overlapping red boxes linked by black lines) was predicted by both GENSCAN (open red boxes linked by red lines) and FGENESH (open blue boxes linked by blue lines). A BLASTX homology (shown as blue box) was also observed. Primers designed to the first exon (shown as red arrows) brought up no positives when tested against all available cDNA libraries by PCR (data not shown).*

4.8 Appendix

Table 4.5: *Link information as described in Figure 4.9*

Link	Accession	Clone Name	Link	Accession	Clone Name
Link_bA810O3	AL355812	RP11-810O3	Link_dJ170D19	AC004822	RP1-170D19
	AL121878	RP6-204F4		AC007025	RP4-673N16
	AL445164	RP4-736G20	Link_dJ29I24	AC007022	RP1-29I24
	AL589786	RP11-161I19		AL589677	RP11-12P4
	AL109751	RP1-237H22	Link_bA320L24	AL441887	RP11-320L24
	AL135921	RP4-682C13		AC004835	RP4-555N2
	AL049591	RP5-878I13		AC004973	RP5-1139I1
	AC003983	RP1-93I3		AC004000	RP3-404F18
	AL589842	RP11-268A15		AC005190	RP5-1152D16
	AC005000	RP1-241P17		AC004913	RP5-876A24
Link_bB377L5	AL513265	RP13-377L5		AL355348	RP13-163A20
	AL121879	RP5-961O8		AC005052	38K21
	AL356314	RP5-858F2		AC002477	RP3-327A19
	AL034411	RP4-808P6		AC005023	GS1-421I3
	Z96810	RP3-452H17		AC006147	RP4-755D9
Link_dJ2C6	AL590156	RP13-420K18		AC002086	RP3-525N14
	AC002071	RP1-2C6		AL512286	RP11-45J1
	AL590157	RP13-420K18		AC006962	RP6-52J4
	AC073306	RP4-564D24		AC002476	RP1-318C15
	AC004823	RP6-172A13		AL451005	RP11-92B10
	AC005002	RP3-378P9		AC011890	RP4-655L22
	AC004959	RP5-1098A23		AC008162	RP1-321E8
	AC006395	RP3-394H4	Link_dJ296G17	AC006144	RP1-296G17
	AL591505	RP11-67I12		AC002377	RP1-222H5
	AL031074	RP1-143G3		AL450488	RP11-161O12
	AC006975	RP5-1026B21		AC007486	RP5-1015P16
	AC008059	RP3-409F10	Link_dJ368G6	AC007074	RP3-368G6
	AL445246	RP11-247H9		AC006143	RP1-74M20
	AC006963	RP1-278D1		AC006314	RP1-314H24
	AC006968	RP4-649M7	Link_bB16D10	AL357562	RP13-16D10
	AL590114	RP11-318H18		AL589847	RP13-477D10
	AC003012	RP1-169K13		AL513487	RP13-63I15
	AC007088	RP6-39H21	Link_bA438H17	AL359956	RP11-438H17
	AL391358	RP11-197B12		AL109800	RP6-64P14
	AL391474	RP13-115H14		Z83848	RP1-57A13
	AL391830	RP13-318F20		AL035426	RP3-370N13
	AL391803	RP13-25C19		Z82899	RP1-181N1
	AC007021	RP6-155F9		AL356213	RP5-1171F9
	AL391237	RP13-125M24			
	AL391280	RP13-128O4			
	AL589824	RP11-76G11			
	AL606485	RP11-370H3			
	AC006965	RP4-562J12			

Table 4.6: *Information of pseudogenes*

Pseudogene	Description
bA320L24.CX.1	similar to ADP-ribosylation factor-like protein 5
bB192B19.CX.1	similar to translationally controlled tumor protein
dJ29I24.CX.1	similar to 60S Ribosomal protein
dA204F4.CX.2	similar to zinc finger protein
dA204F4.CX.3	similar to YES-associated protein
dJ1189B24.1	similar to NADH-Ubiquinone Oxidoreductase MLRQ subunit
dJ1189B24.2	similar to Tubulin Beta
dJ1189B24.3	similar to Proto-oncogene Tyrosine-protein Kinase
dJ169K13.CX.2	similar to Tubulin Beta
dJ169K13.CX.4	similar to Ribosomal Protein L12
dJ170D19.CX.3	similar to Heat shock cognate 70
dJ222H5.CX.4	similar to mitochondrial heat shock protein 70 (hsp70)
dJ237H22.CX.2	similar to activator of apoptosis Hrk (HRK)
dJ241P17.CX.1	similar to arginosuccinate synthetase
dJ241P17.CX.2	similar to elongation factor-1-gamma
dJ321E8.CX.1	similar to cell cycle protein p38-2G4 homolog (hG4-1)
dJ378P9.CX.1	similar to transcription factor, CA150
dJ525N14.CX.2	similar to elongation factor Tu family
dJ555N2.CX.2	similar to heterogeneous nuclear ribonucleoprotein
dJ562J12.CX.1	similar to aflatoxin aldehyde reductase AFAR

Table 4.7: *STSs used for mutation screening of MRX23 patients*

STS name	Primer 1	Primer 2	Size (bp)	AT (°C)	Gene	Exon
stdJ808P6.4	GAGCTTCTCTTCATAAATG	GTGGAGCACAAGGAACAG	298	55	hATBo+	1
stdJ808P6.8	TGTTGCTCTATGGATTTG	ATGCCTTCCTTCAACTCG	354	55	hATBo+	2
stdJ808P6.5	TAGAAAGCAGTGAACTTTAG	CAGTGAAGGTAGCTATATG	361	55	hATBo+	3
stdJ808P6.6	TGAGATACAGCTTTTTTATG	TGCTTTCACCAGTGACCTTTG	366	55	hATBo+	4
stdJ808P6.13	GATGCTGAATGTACATAGC	CACATTGCTGACTATGAGC	517	53	hATBo+	5
stdJ808P6.14	CTATCTGTGCCCTTCGTATTG	GCTTGATATTGAACTACCATG	376	55	hATBo+	6
stdJ808P6.10	TAACTTTGGTATATCATCAG	TTGCAAGCTATTACATTATG	388	55	hATBo+	7
std452H17.12	GTAGAAGGGTGACAATGATG	CTATTGGAGTTTCATAAGTG	494	55	hATBo+	8
stdJ452H17.3	AGATTTTTCTAATATCTTATG	CTTTCAGAAAGATCATTCTG	343	55	hATBo+	9
stdJ452H17.4	TGAATTCTGTGATTAACAG	ACCTGGACTTGTCACTAAG	293	55	hATBo+	10
std452H17.13	GAAACTAAGGAGCATATG	GTTGTGCAGTATATTGTAC	343	53	hATBo+	11
stdJ452H17.6	ACAGAAAGATAATTGATG	TTGCCTTTTGTCTTCAATG	276	55	hATBo+	12
stdJ452H17.7	TGATAAATCACATCTGAG	AGGTATAGAAGTAGCCAAG	401	55	hATBo+	13
stdJ452H17.8	CAGTTCAATATTTGCTTG	ACATGGCTGAGAATTAAGA	405	55	hATBo+	14
stdJ93I3.4	GTTGATGTGACAGGCTCG	TGAGCTTAACCGAGATGC	315	55	T-plastin	1
stbA268A15.1	AGATGAGAACTTAGCAAG	AGAGAAATAACTTTGAGAC	221	55	T-plastin	2
stbA268A15.2	GTGAGCTTATGAACTGAAC	GATATTCCAGCAGCTAAAAG	433	60	T-plastin	3
stbA268A15.3	GAGTTCAATGCATGTAGC	GCCCGTCCTTGACATTAC	448	60	T-plastin	4
stbA268A15.4	ACCACTGTGTTGCATCCAAG	GATTCATGGACAGACCTAG	460	60	T-plastin	5
stdJ241P17.1	GAGTACATGAAAGAGATG	GAACAGGTCCTCAAACAG	257	55	T-plastin	6
stdJ241P17.2	GAAAGGTCAAGAAGCAAGTG	GCACGAAAGTCTGCATGAC	276	55	T-plastin	7
stdJ241P17.3	GACTGAATGAACTTGGCATG	CTTGGTGATACAGTGTTAGG	313	55	T-plastin	8
stdJ241P17.4	CATGGGACAATAGGATAC	GCCAACTCTACTTCATACG	262	55	T-plastin	9
stdJ241P17.5	GTAGAACTGTATACCCAG	GCCATCTACTTCTTGTAG	403	53	T-plastin	10
stdJ241P17.6	GTGTATTGGCACTATATGC	CATCCATTCATGACATTCG	201	55	T-plastin	11
stdJ241P17.7	CTTGTTTGACAATGTAGTG	CTCTAACAAATATATACAGC	239	55	T-plastin	12
stdJ241P17.8	CATTTACTCTTGTGCCTTTG	GTGTAGTTATCGACATATC	255	55	T-plastin	13
stdJ241P17.9	CTCATAAAGTAGATGGTGAC	GTAAAGAATTGTGCCATTAG	291	55	T-plastin	14

stdJ241P17.10	GGCTTCTTTGTGAGTGAG	CTATTAGCAGTCTCCCTTAC	292	55	T-plastin	15
stdJ241P17.11	GTGTCCTTAACTGACAAG	GCCAAGAGTTCCTTAAGC	282	55	T-plastin	16
stdJ241P17.12	CTTCTGCAGCTCCTGGTG	CGGTAGTAGTCAGGATGTG	488	55	SMT3B	1

Table 4.8: *Information on cDNA sequencing projects*

cDNA collection	**Web Address**
Mammalian Gene Collection (NIH) (MGC)	http://mgc.ncbi.nih.gov/
NEDO database (Kazusa DNA Research Institute) (FLJ)	http://www.kazusa.or.jp/NEDO/
HUGE database (Kazusa DNA Research Institute) (KIAA)	http://www.kazusa.or.jp/huge/
The German Human cDNA project (DKFZ)	http://mips2.gsf.de/proj/cDNA/
The Riken Mouse collection	http://genome.rtc.riken.go.jp/
Genoscope	http://www.genoscope.cns.fr/
EMBL (dBEST, vert_RNA)	http://www.ebi.ac.uk/embl/

Chapter 5

Comparative Sequence Analysis

Between Human and Mouse

5.1 Introduction

Humans and mice diverged from a common ancestor approximately 70 million years ago and have a comparable genome size (O'Brien, S. J., *et al.*, 1999). Comparison of orthologous genes in human and mouse and their function has shown that sequence similarity across much of the coding regions of genes and some of the regulatory elements that control them has been maintained since the split from a common ancestor. Some of the early evidence of conservation between human and mouse came from comparative analysis of 100 kb of human and mouse T-cell receptor DNA (Koop, B. F., *et al.*, 1994). More recently, regions of conservation have been identified upstream of the SCL gene in human, mouse and chicken, and were later shown to be associated with active regulatory regions (Gottgens, B., *et al.*, 2001).

The striking sequence similarity between human and mouse in specific genomic regions arises because functionally conserved features between genomes tend to be conserved at the sequence levels. This allows for inferences to be made about one organism using information determined in the other. Comparative sequence analysis is therefore a powerful tool for aiding both human gene identification and understanding the function and control of genes. As discussed in the previous chapter, the function of only a small proportion of human genes identified to date has been experimentally determined. The identification of the orthologous genes between human and mouse will enable function of the human counterpart to be inferred based on the investigation into the function of the orthologue in mouse.

The evolution of mammalian sex chromosomes is characterised by the loss of genes from the Y chromosome by mutation. This, in turn, has led to the development of X inactivation in females in order to achieve dosage compensation for X-linked gene products. Ohno's law states that due to dosage compensation in females, it is thought there is selective pressure to maintain dosage-dependent genes on the X chromosome (Ohno, S., 1967). In agreement with Ohno's suggestion, X-linkage of genes is generally maintained in the eutherian mammals, which is in contrast to what has been observed for the autosomes. The human X chromosome is currently represented by nine syntenic blocks all positioned on the mouse X chromosome (see Figure 5.1). In contrast, human chromosome 6, a similarly sized chromosome to the human X chromosome is represented by at least eight syntenic blocks in the mouse, but present on seven different mouse chromosomes (data taken from http://www.ncbi.nlm.nih.gov/Omim/Homology/human6.html). The availability of orientated X chromosome sequence in both human and mouse allows for a refinement of the synteny map, enabling the precise order and transcriptional orientation of genes to be studied.

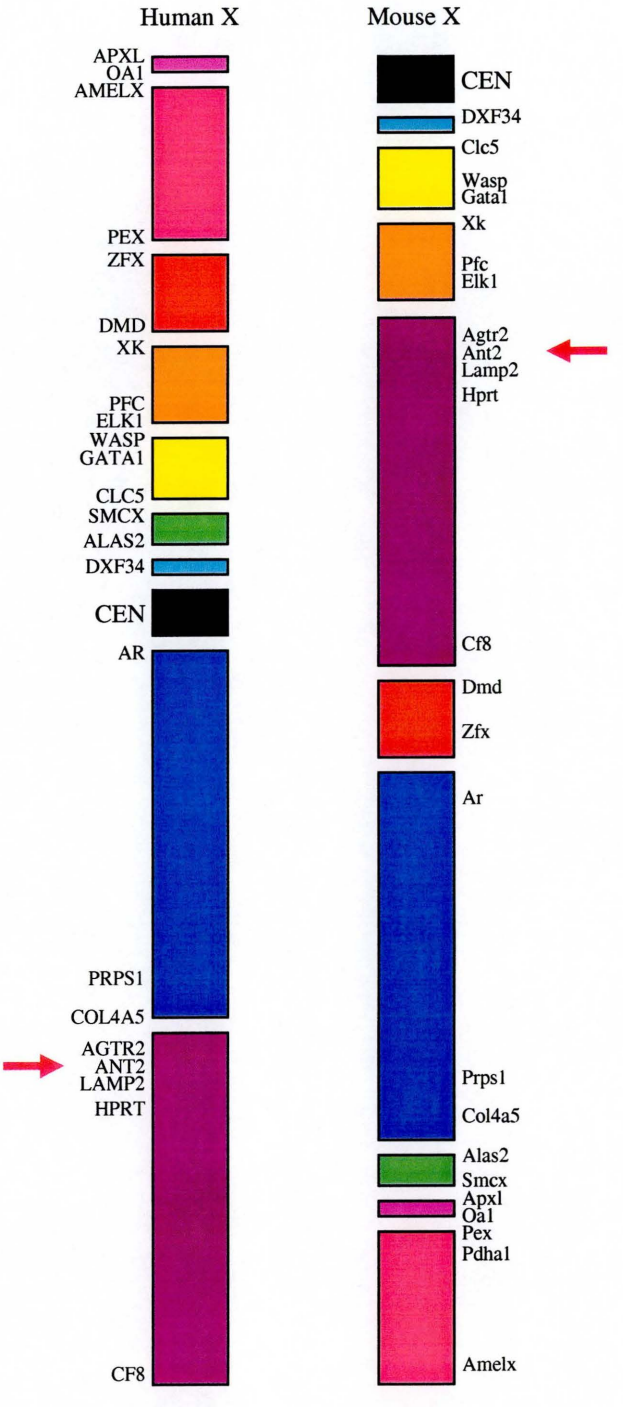
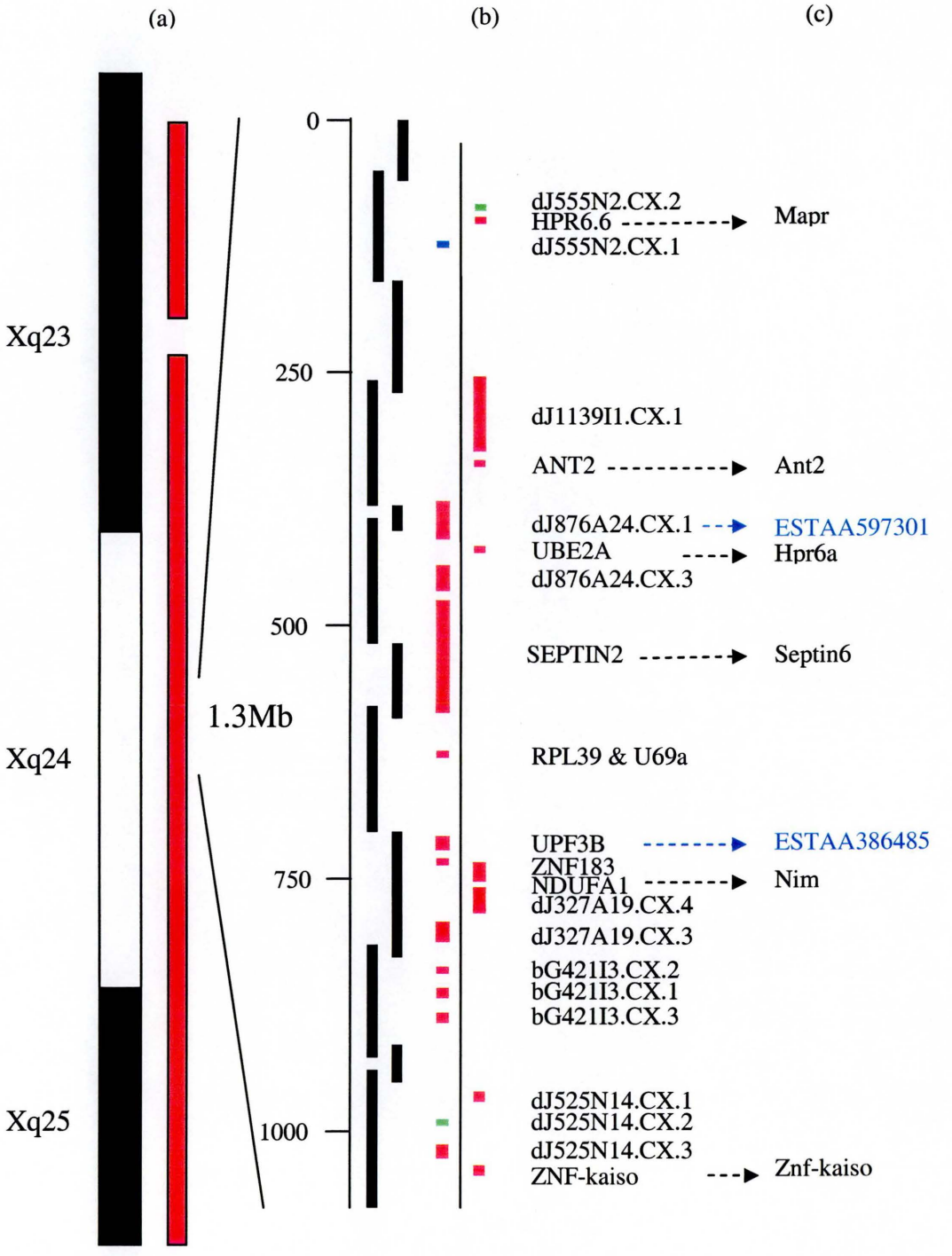


Figure 5.1: *A schematic representation of the syntenic relationship between human and mouse (reproduced from Boyd, Y. et al., 1998). The human X chromosome (left) is divided into 9 syntenic blocks when compared to the mouse X chromosome (right). Syntenic blocks are shown in the same colour. A subset of the markers known to map to each block is also shown. The position of the region studied in this chapter, estimated from the position of mouse Ant2 gene, is shown as a red arrow.*

In this study, a contiguous segment of finished sequence in human Xq24, between HPR6.6 and ZNF-kaiso was chosen for comparative analysis with the mouse. The region was chosen because of the advanced state of the human sequencing and annotation at the time (as discussed in the previous chapter). The 1.3 Mb region contains twenty genes, of which nineteen have been confirmed by cDNA sequence and one remains predicted, and one pseudogene (see Figure 5.2). The syntenic region in mouse is thought to be located in the proximal section of the fifth syntenic block, based on the position of one known orthologous gene, Ant2 (arrowed in Figure 5.1). The aim of the work contained within this chapter was to investigate the usefulness of mouse sequence for annotating human genes, and generate a detailed comparative map of orthologous genes in the region.

Figure 5.2: *(see over) Summary of the region for comparative analysis. (a) The position of the region of interest in relation to the transcript map described in the previous chapter. (b) The minimum set of clones (denoted as vertical black bars) and the genes identified (genes in red, predicted genes in blue, pseudogenes in green). Genes on the left of the thin vertical black line are transcribed on the minus strand, genes on the right are transcribed on the plus strand. A scale is shown in kilobases. (c) The known orthologous genes (shown in black) or ESTs (shown in blue) used for bacterial clone isolation. Arrows link orthologous sequences.*

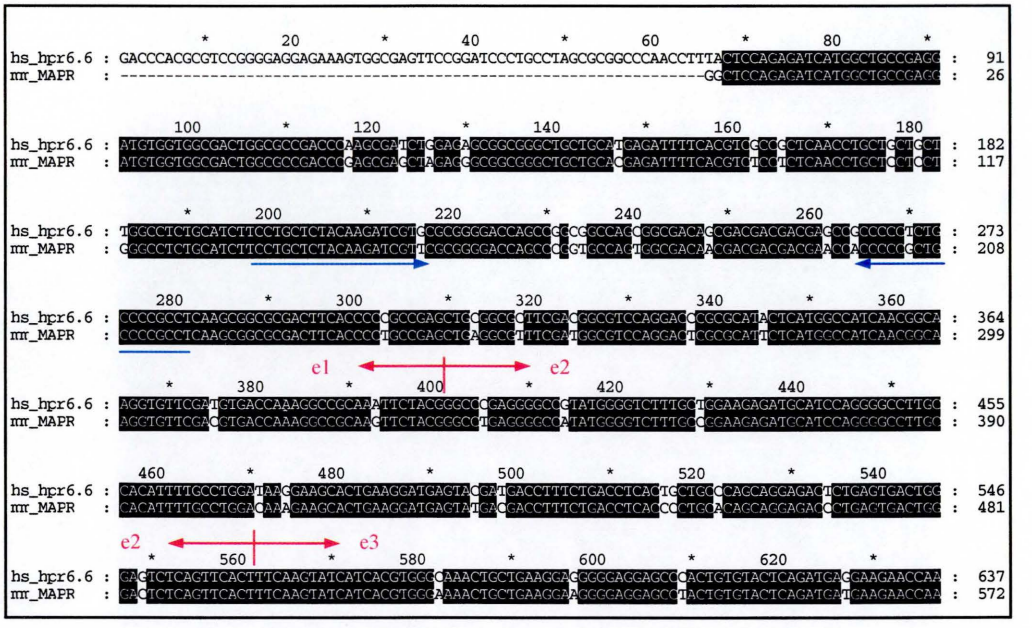

(a)
(b)
(c)
Xq23
Xq24
Xq25
1.3Mb
0
250
500
750
1000
dJ555N2.CX.2
HPR6.6
dJ555N2.CX.1
Mapr
dJ1139I1.CX.1
ANT2
Ant2
dJ876A24.CX.1
ESTAA597301
UBE2A
Hpr6a
dJ876A24.CX.3
SEPTIN2
Septin6
RPL39 & U69a
UPF3B
ESTAA386485
ZNF183
NDUFA1
Nim
dJ327A19.CX.4
dJ327A19.CX.3
bG421I3.CX.2
bG421I3.CX.1
bG421I3.CX.3
dJ525N14.CX.1
dJ525N14.CX.2
dJ525N14.CX.3
ZNF-kaiso
Znf-kaiso

RESULTS

5.2 Construction of bacterial clone contig

The identification of eight mouse-specific sequences expected to map to the syntenic region in mouse provided the basis for the construction of the bacterial clone contig (see Figure 5.2c). These consisted of six mouse mRNAs known to be orthologous to human genes in the region and two mouse ESTs that were greater than 90% identical at the nucleotide level to human genes. For each orthologous pair, the two sequences were aligned in order to identify the most likely positions of the introns in the mouse gene, based on the positions of the human introns. Examples of the alignments can be seen in Figure 5.3. For each orthologous pair, a PCR assay was then designed within a single exon of the mouse sequence, and used for mouse genomic bacterial clone isolation (see Figure 5.4).

(a)



(b)

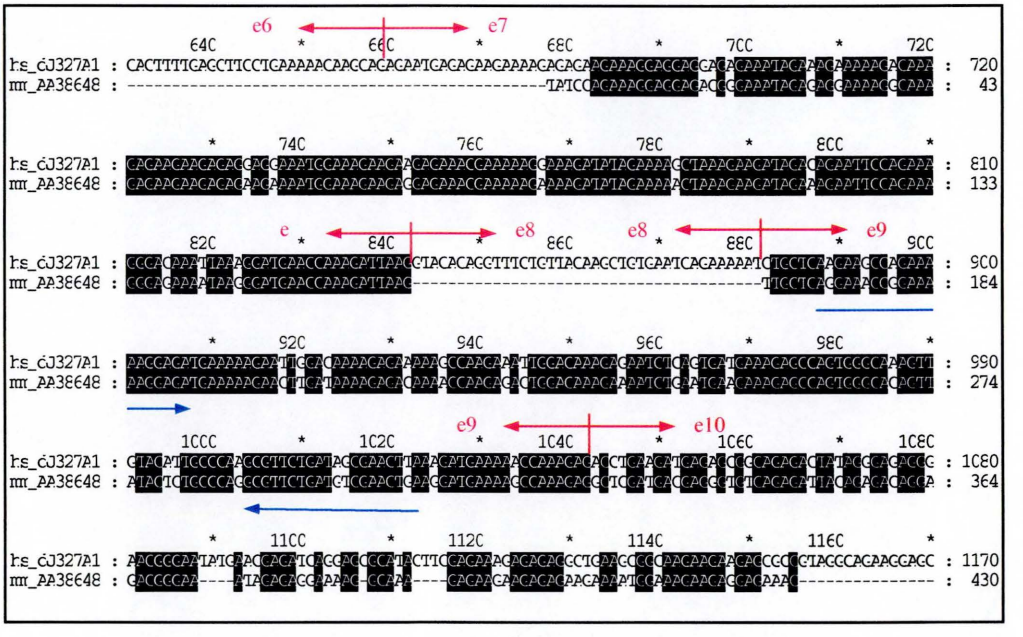


Figure 5.3: *Examples of alignment between human and mouse orthologues. Sequences are aligned using CLUSTALW and visualised in GENEDOC. The boundaries between exons are shown as red arrows and the positions of primers for the STS are shown in blue. (a) An alignment between HPR6.6 (human) and MAPR (mouse). (b) An alignment between part of UPF3B and the EST AA386485.*

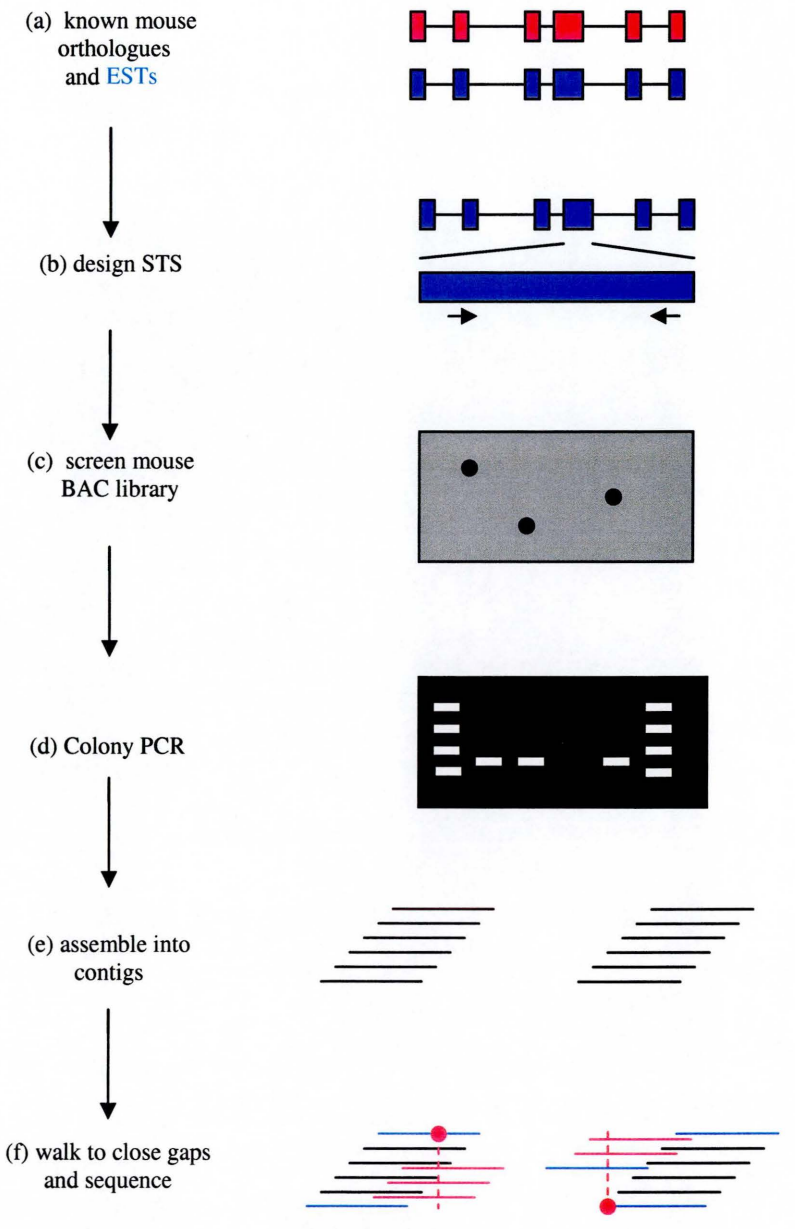
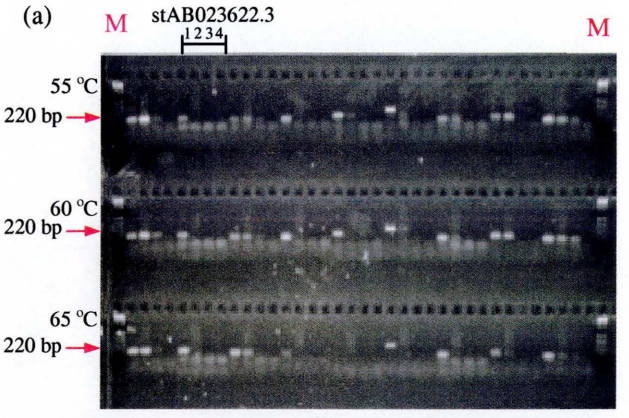
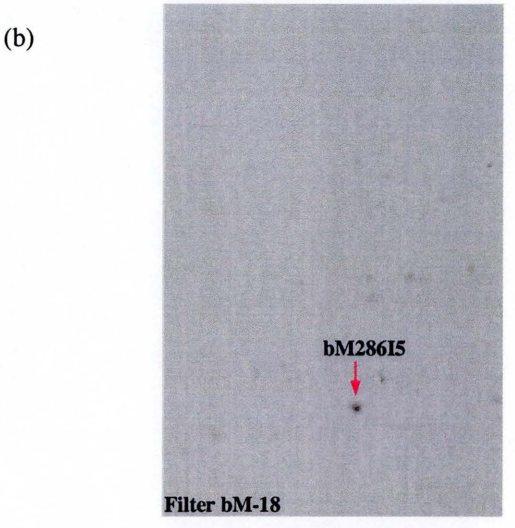


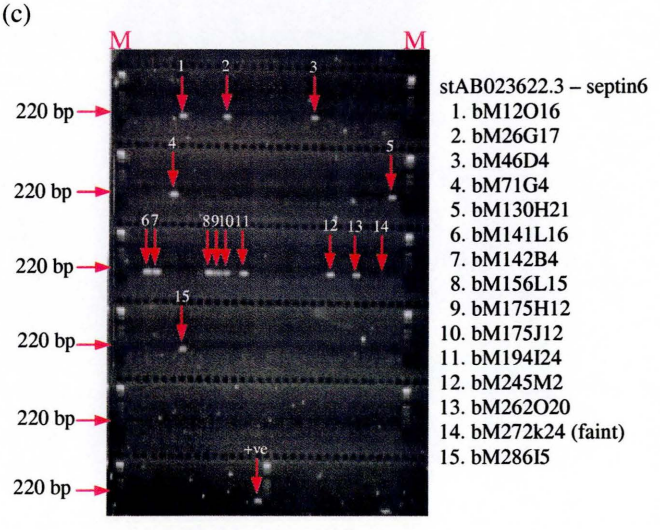
Figure 5.4: *Strategy for contig construction (a) Orthologous genes are aligned (human shown in red and mouse shown in blue) and (b) STSs designed within a single exon (primers shown as black arrows). (c) A pool of STSs is hybridised to a filter (shown as a grey rectangle) and positive BACs (black spots) are identified. (d) BACs positive for each STS are (shown as white rectangles) and (e) clones are assembled into contigs by fingerprinting (horizontal black lines). (f) New STSs (red dot) are identified at the ends of contigs for contig extension and a minimum set of clones identified for sequencing (shown as blue horizontal lines).*

The eight PCR assays, designed to the orthologous mouse sequences, were pooled and hybridised to a gridded array of BAC clones (RPCI-23, see Section 2.8.2) (see Figure 5.5). A total of 45 BACs were identified and positive clones for each PCR assay were confirmed using colony PCR. All BACs identified within the region were assembled into contigs using *Hind* III fingerprinting (see Section 2.12.3). At this stage, there were 2 contigs covering 1.1 Mb, estimated using fingerprint band sizes (see Section 2.23.2). A section of the contig showing the integration of clones positive for stAB023622.3, designed within an exon of the mouse septin6 gene, is shown in Figure 5.6.

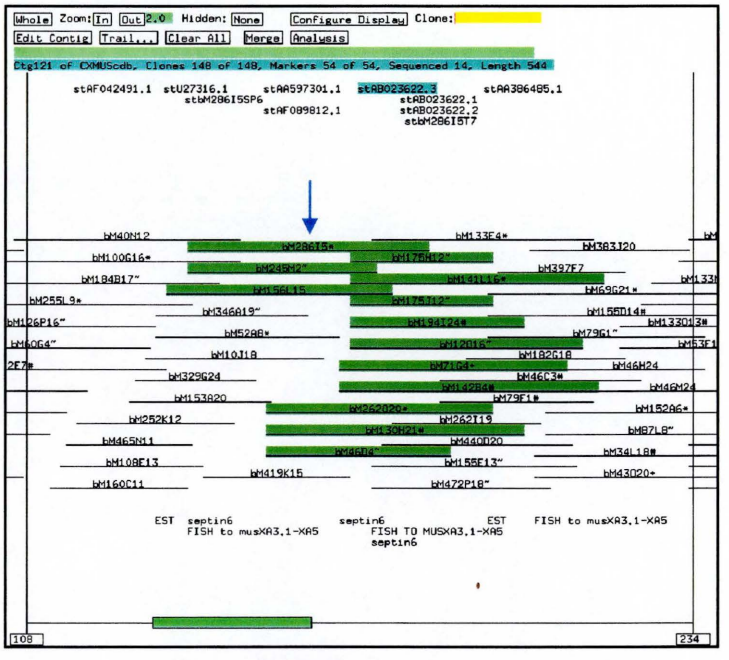
Figure 5.5: *(see over) BAC clone isolation with mouse-specific STSs. (a) Nine STSs designed from sequence generated to nine mouse-specific sequences were tested for their ability to amplify unique sequence in mouse genomic DNA at three different temperatures of the PCR. Reactions included mouse genomic DNA (1), human genomic DNA (2), human X-chromosome hybrid (3), and T0.1E (4). M = marker (b) The product of amplification of mouse genomic DNA using stAB023622.3 designed to the mouse Septin2 gene was labelled, along with eight other products, pooled and used as a hybridisation probe to screen gridded filters containing BAC clones from RPCI-23. The filter shown, bM-18, has one positive clone as marked. (c) Positive clones detected on all filters were streaked and individual colonies tested against stAB023622.3 Positive clones are listed to the side.*


(a)
M
stAB023622.3
1 2 3 4
M
55 °C
220 bp
60 °C
220 bp
65 °C
220 bp


(b)
bM286I5
Filter bM-18


(c)
M
M
1
2
3
4
5
6 7
8 9 10 11
12
13
14
15
+ve
220 bp
220 bp
220 bp
220 bp
220 bp
220 bp
stAB023622.3 – septin6
1. bM12O16
2. bM26G17
3. bM46D4
4. bM71G4
5. bM130H21
6. bM141L16
7. bM142B4
8. bM156L15
9. bM175H12
10. bM175J12
11. bM194I24
12. bM245M2
13. bM262O20
14. bM272k24 (faint)
15. bM286I5

(a)



(b)

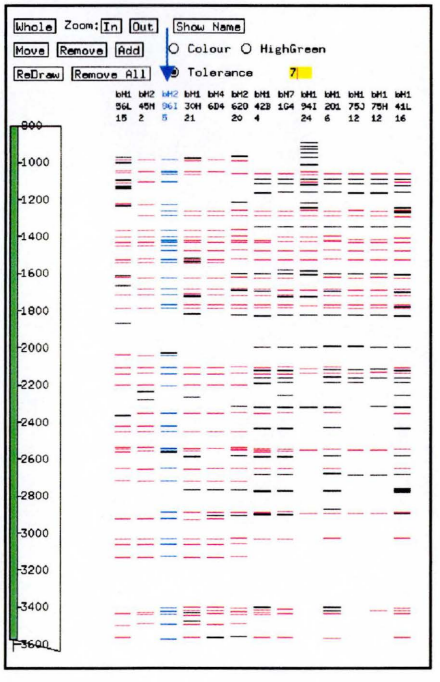


Figure 5.6: *Contig construction by fingerprinting. A section of the contig from FPC, showing the positive clones (highlighted in green) for stAB023622.3 (highlighted in blue) and (b) their fingerprints. The fingerprints of the third positive clone (counting from the left) are shown in blue, and bands with comparable migration distances for the clones in the other lanes are shown in red.*

At this time, sequence was being generated at the ends of all BACs in the RPCI-23 library (http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.html). In an attempt to close the gap between the two contigs, 13 STSs were designed to the available end sequence from 9 BACs that were positioned close to the ends of each contig. A pooled probe containing the thirteen STSs (see Table 2.7) was hybridised to the gridded array of RPCI-23 BACs, and the positive clones confirmed by colony PCR. The 53 newly identified BACs were fingerprinted and incorporated into the existing contigs. At this stage, there were still two contigs but these had been extended to cover 1.85 Mb. In a further attempt to close the remaining gap, a second pooled probe containing two novel STSs, stbM206F21SP6 and stbM202F23SP6, one from each end of the contigs, failed to identify any new clones when hybridised to the two mouse genomic libraries available at the time (RPCI-23). It was concluded that there were no clones in the available mouse clone libraries that bridged the gap between the two contigs.

In summary, a region covering 1.9 Mb has been covered in two bacterial clone contigs of 1.1 Mb and 0.75 Mb respectively (see Figure 5.7a) and the gap has been sized at approximately 50 kb using fibre-fish (carried out by Pawandeep Dhami) (see Figure 5.7b). A minimum tiling set of seven clones were chosen for sequencing and there are two contiguous segments of finished sequence available covering 714 kb and 193 kb (December 2001) (see Figure 5.7c). A further three clones are available as draft sequence. Four of the clones identified for genomic sequencing (bM100G16, bM38B5, bM43O20 and bM322E15) were localised to XA3.1-XA5 by FISH onto metaphase spreads of mouse chromosomes (carried out by Sheila Clegg), which includes the region syntenic to human Xq24.

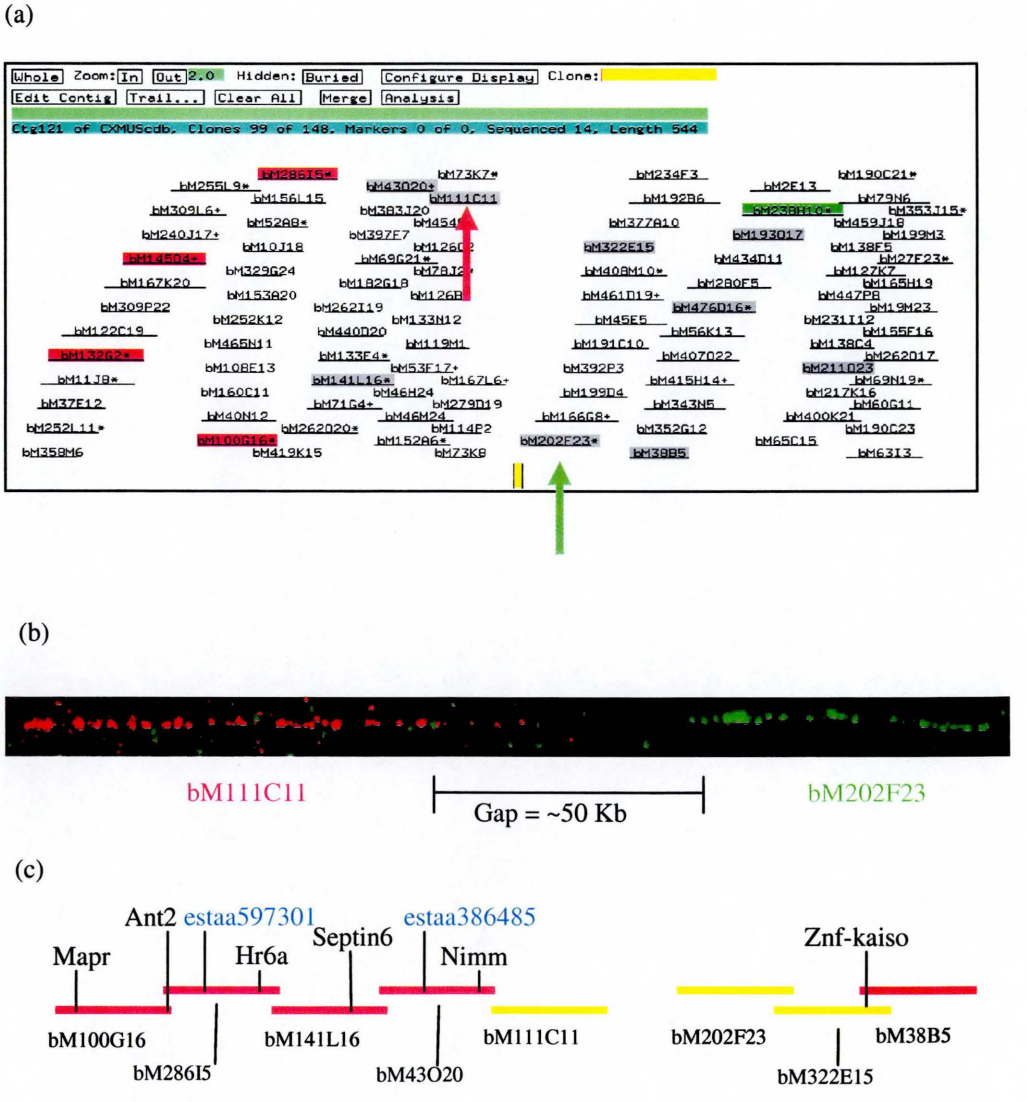
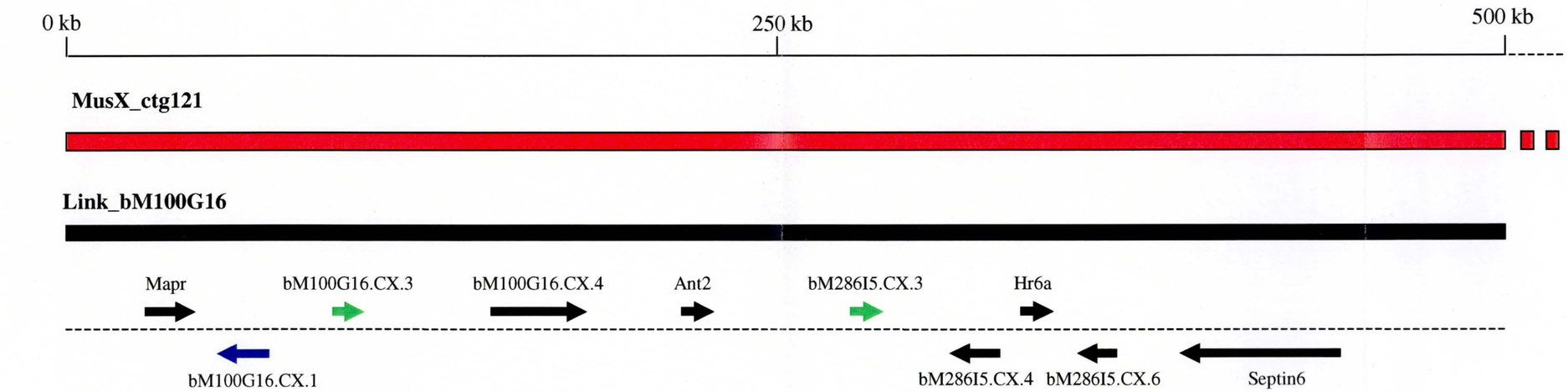
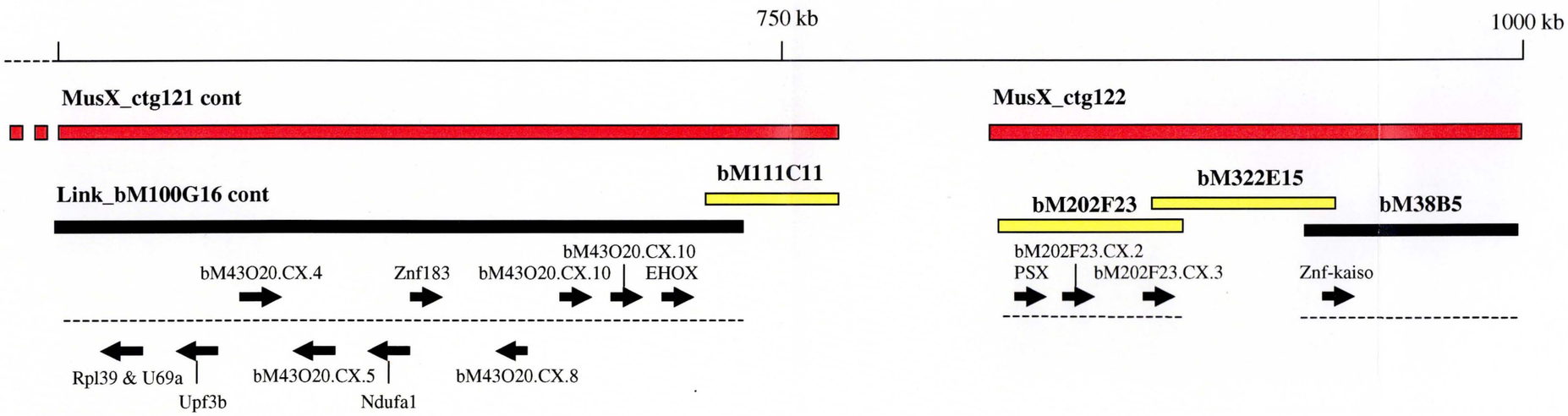


Figure 5.7: *Summary of the mapping. (a) The final contig viewed in FPC. The minimum set of clones identified for sequencing are highlighted (red = finished, grey = draft sequence, green = picked for sequence). Clones used to size the gap are indicated by a red arrow (bM111C11) and a green arrow (bM202F23). (b) The fibre-fish image showing the size of the gap with respect to the length of signal for each clone (approximately 1/3 the length of bM111C11, suggesting the size of the gap is approximately 50 kb). (c) The minimum set of clones sequenced (red = finished, yellow = draft shotgun) and the positions of the mouse-specific genes and ESTs used during the construction of the contig.*

All available genomic sequence data has been analysed as described in the previous chapter (Section 4.2). The analysis used a combination of computational gene prediction and similarity searches, matching genomic sequence to all known DNA and protein sequences. The region was found to contain a total of twenty-four genes, twenty-one genes confirmed by previously available cDNA sequence, one predicted gene and two pseudogenes (see Figure 5.8). No attempts have been made to identify confirmatory cDNA sequences for the predicted gene due the lack of availability of mouse cDNA resources at the time.

Figure 5.8: *(see over) Summary of the gene map constructed in mouse. The red bars indicate the status of the contigs and the black bars indicate the extent of finished sequence. Each link represents a series of individual clones (see appendix to this chapter). Yellow bars indicate clones for which draft sequence was available as of December 2001. A scale is given in kilobase pairs (kb). Approved names are given for known genes. Genes are indicated by arrows (black – complete, blue – predicted, green – pseudogene), the direction of each arrow reflects the direction of transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.*


0 kb
250 kb
500 kb
MusX_ctg121
Link_bM100G16
Mapr
bM100G16.CX.3
bM100G16.CX.4
Ant2
bM286I5.CX.3
Hr6a
bM100G16.CX.1
bM286I5.CX.4
bM286I5.CX.6
Septin6


750 kb
1000 kb
MusX_ctg121 cont
Link_bM100G16 cont
bM111C11
bM43O20.CX.4
Znf183
bM43O20.CX.10
bM43O20.CX.10
EHOX
Rpl39 & U69a
Upf3b
bM43O20.CX.5
Ndufa1
bM43O20.CX.8
MusX_ctg122
bM322E15
bM202F23
bM38B5
bM202F23.CX.2
PSX
bM202F23.CX.3
Znf-kaiso

5.3 Identification of orthologous genes in the region

Orthologous genes are defined as being homologous genes in different organisms derived from the same gene during speciation (Postlethwait, J. H., *et al.*, 1998). It can be very difficult to determine the true relationship of two genes from different species as both species have been undergoing independent evolution since they diverged from the common ancestor. If a gene has duplicated within a species since the divergence to create a pair of paralogues, sequence similarity alone is not sufficient to be certain that two similar genes in different species are derived from a single gene in a common ancestor. However, a number of features from each gene can be compared in order to ascertain whether two genes are likely to be derived from the same common ancestral gene. These include:

1. Similarity at the nucleotide and amino acid level.
2. Similarity of exon and intron structure.
3. Position with respect to neighbouring genes (and correspondence of identity of neighbouring genes, i.e. synteny).
4. Function if known.
5. Lack of other similarly matching sequence in rest of either genome.

Comparison of the genes found between HPR6.6 and ZNF-Kaiso in human with those identified in mouse using the criteria described above, reveals a total of sixteen pairs of genes which appear to be orthologous (see Figure 5.9 and Table 5.1). Six of these had been previously identified and were used in the construction of the bacterial clone contigs and the remaining nine have been determined in this study.

Each orthologous pair shows a high level of similarity at both the nucleotide and amino acid level and good conservation of exon structure. An example of an orthologous pair is shown in Figure 5.10.

Figure 5.9: *(see over) Comparative analysis of the region in human (on the left) and mouse (on the right). The position of genes (red = gene confirmed by cDNA, blue = predicted gene, green = pseudogene) and their direction of transcription (minus strand on the left of vertical line, plus strand on the right) are shown. The names of the genes used during the construction of the contig are shown in blue. Segment 1 (indicated by a vertical purple line) in human and mouse shows a high level of synteny. Segment 2 (indicated by a vertical green line) shows the extent of the inversion of the four genes. Segment 3 (indicated by the veritcal gold line) appears to contain apparent non-orthologous genes. The genes predicted by INTERPRO to contain a homeobox domain are indicated in bold. The genes (ZNF-Kaiso and Znf-kaiso) in segment 4 (blue line) are orthologous.*

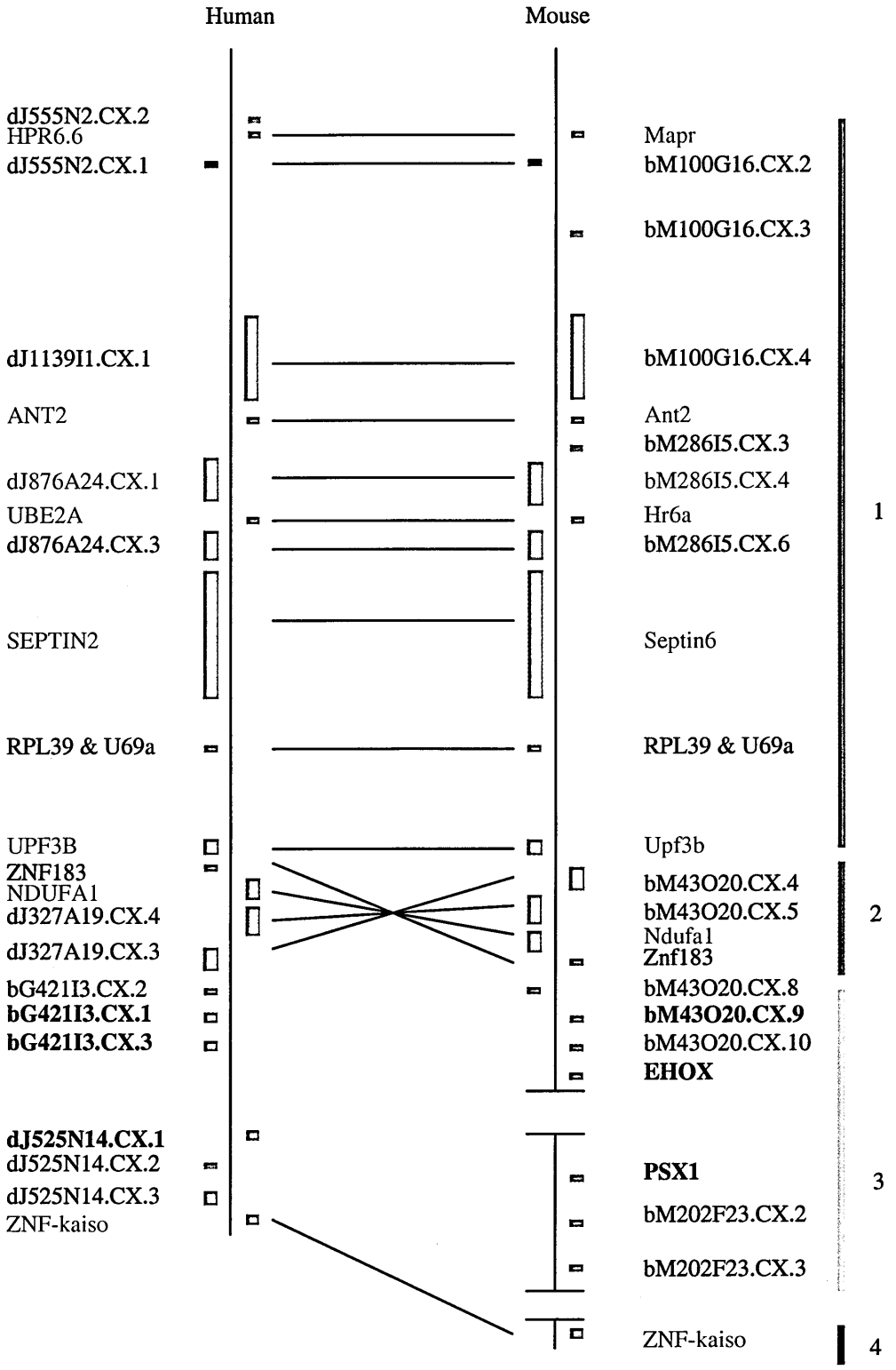
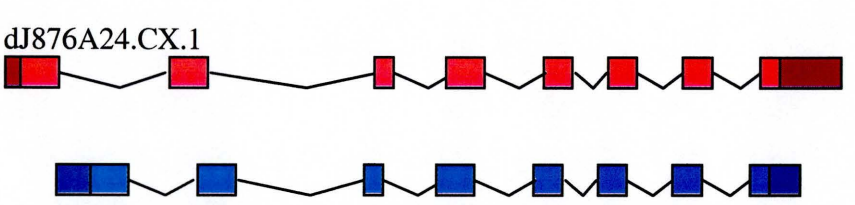

Human
Mouse
dJ555N2.CX.2
HPR6.6
dJ555N2.CX.1
dJ1139I1.CX.1
ANT2
dJ876A24.CX.1
UBE2A
dJ876A24.CX.3
SEPTIN2
RPL39 & U69a
UPF3B
ZNF183
NDUFA1
dJ327A19.CX.4
dJ327A19.CX.3
bG421I3.CX.2
bG421I3.CX.1
bG421I3.CX.3
dJ525N14.CX.1
dJ525N14.CX.2
dJ525N14.CX.3
ZNF-kaiso
Mapr
bM100G16.CX.2
bM100G16.CX.3
bM100G16.CX.4
Ant2
bM286I5.CX.3
bM286I5.CX.4
Hr6a
bM286I5.CX.6
Septin6
RPL39 & U69a
Upf3b
bM43O20.CX.4
bM43O20.CX.5
Ndufa1
Znf183
bM43O20.CX.8
bM43O20.CX.9
bM43O20.CX.10
EHOX
PSX1
bM202F23.CX.2
bM202F23.CX.3
ZNF-kaiso
1
2
3
4

Table 5.1: *Comparison of orthologous genes*

Human Gene	Mouse Gene	Exon no. human	Exon no. mouse	% identity protein
HPR6.6	Mapr	3	3	98
dJ1139I1.CX.1	bM100G16.CX.4	5	5	77
ANT2	Ant2	4	4	98
dJ555N2.CX.1	bM100G16.CX.2	4	4	64
dJ876A24.CX.1	bM286I5.CX.4	7	7	99
UBE2A	Hr6a	6	6	100
dJ876A24.CX.3	bM286I5.CX.6	2	2	96
SEPTIN2	Septin6	8	8	98
RPL39	Rpl39	3	3	100
UPF3B	Upf3b	11	11	93
ZNF183	Znf183	1	1	90
NDUFA1	Ndufa1	3	3	94
dJ327A19.CX.4	bM43O20.CX.5	5	5	61
dJ327A19.CX.3	bM43O20.CX.4	9	9	94
ZNF-KAISO	Znf-kaiso	2	2	92

A high degree of synteny is apparent between the human and mouse sequence. The proximal portion of the two regions between HPR6.6 and UPF3B in human, and between Mapr and Upf3b in mouse, are exactly conserved in terms of both gene content and gene order (see Figure 5.9, segment 1). The distal portion of the region analysed appears to contain two segments where synteny is disrupted (see Figure 5.9, segments 2 and 3). Segment 2 contains four genes that appear to have undergone an inversion in one of the two species since the divergence from a common ancestor. Analysis of the order of genes in other mammals or vertebrates will enable further investigation into when this inversion took place. There is synteny at the distal end of the region (segment 4) based on the presence of the ZNF-kaiso orthologues.

(a)



(b)

Gene Name	dJ876A24.CX.1	mm_bM286I5.CX.4
5'UTR	46	106
exon 1	124	124
2	118	118
3	42	42
4	139	139
5	90	90
6	93	93
7	63	63
3'UTR	325	187
Total Coding	669	669
Total cDNA size	1040	962
intron 1	4846	267
2	14763	6730
3	971	2430
4	1758	863
5	1084	598
6	1539	1476
Total Intron size	24961	12364

(c)

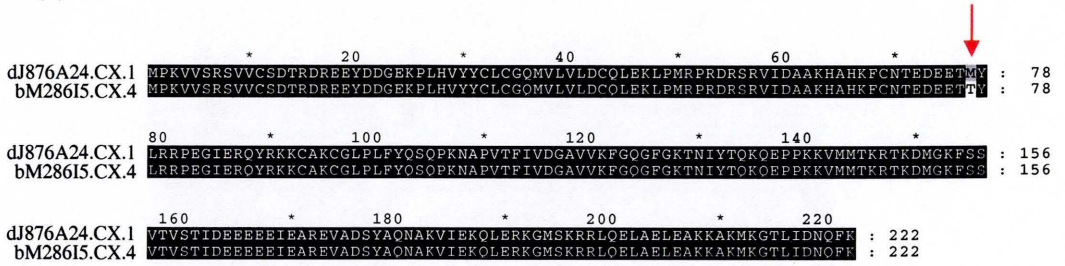


Figure 5.10: *Comparison of a novel orthologous pair of genes. (a) A schematic represention (not to scale) of the gene structure of the human gene (shown in red) and the mouse gene (shown in blue). Exons are shown as boxes, introns shown as 'v' shaped lines. UTR's are shown as darker coloured boxes (b) A comparison of the sizes of exons and introns, total coding sequences and total cDNA size for the two genes. (c) An alignment of the predicted protein sequences of the two genes showing the amino acid difference (indicated with an arrow).*

Segment 3 contains a number of genes that do not appear to be orthologous between the two species. In both human and mouse, segment three contains three genes predicted by INTERPRO (http://www.ebi.ac.uk/INTERPROSCAN) to contain homeobox domains, these are labelled in bold in Figure 5.9 (bG421I3.CX.1, bG421I3.CX.3, dJ525N14.CX.1, in human, and bM43O20.CX.9, EHOX and PSX1 in mouse). Apart from bG421I3.CX.3, which has only three exons, they all have a similar gene structure with four exons and three introns (see Figure 5.11a). An alignment of the predicted protein sequences of the six genes (using CLUSTALW) shows that although there is similarity between all proteins in the region of the homeobox domains, there is no significant similarity for the rest of the alignment (see Figure 5.11b). This is also the case when individual human proteins are aligned with individual mouse proteins (data not shown).

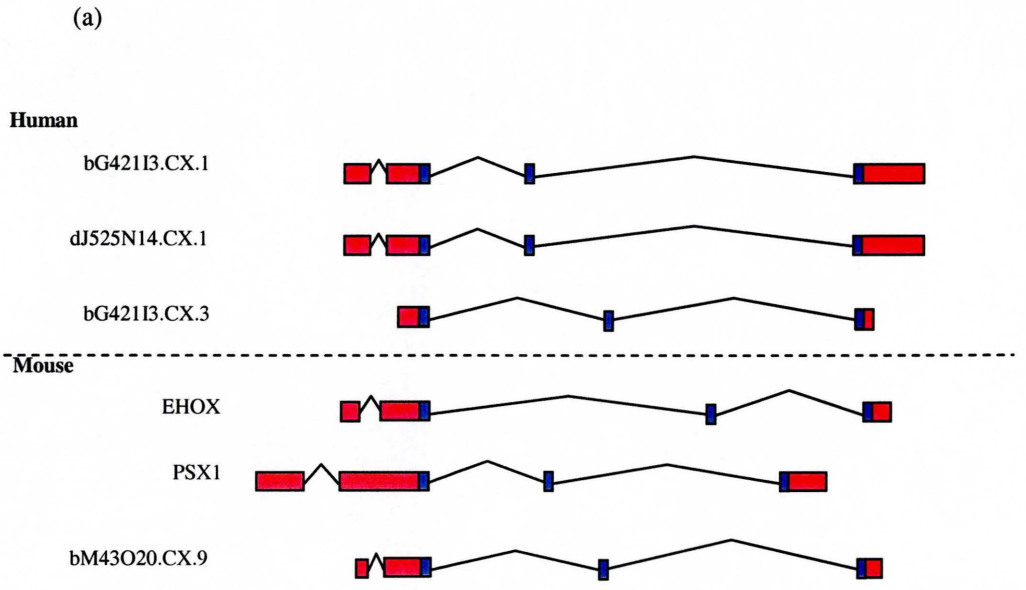


Figure 5.11: *Analysis of the homeobox genes (a) A schematic representation of the six homeobox genes located in segment 3 in human and mouse. Exons are indicated as boxes and introns are indicated as 'v' shaped lines. The region predicted to code for the homeobox domain for each gene is shown in blue.*

(b)

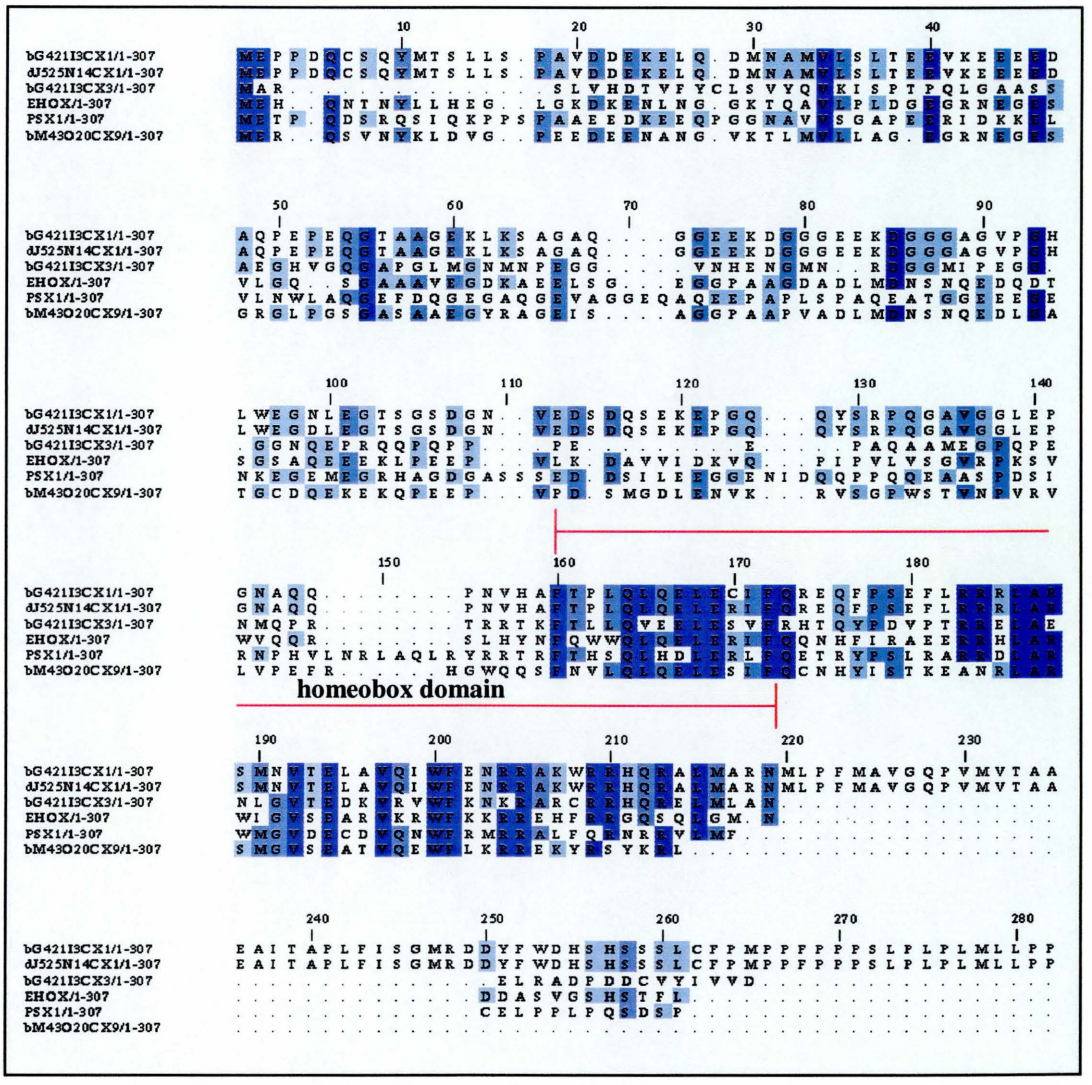


Figure 5.11 cont: *(b) An alignment of the predicted protein sequences of the six genes. Amino acids are given in the one letter code. Amino acids shared by at least five of the six proteins in any one position are shaded in dark blue, and those shared by three of the five proteins are shaded in light blue. A high degree of conservation can be seen in the region predicted to contain the homeobox domain (indicated by the red lines).*

Two of the human genes, dJ525N14.CX.1 and bG421I3.CX.1 are almost identical and are present within a 50 kb inverted repeat (discussed in Section 4.3.4). These appear to have arisen by duplication since human and mouse diverged from a common ancestor, due to the conservation in the positioning of an *Alu* element in each human gene, specifically an *Alu*Sx within the first intron of both genes. As *AluSx* elements arose in human between 32 million and 53 million years ago (Jackson, M. S., *et al.*, 1996), it is likely that the duplication event occurred since humans and mice diverged from a common ancestor, estimated to be approximately 70 million years ago. Therefore, it may have been expected to observe only one orthologue in mouse to these two human genes. However, the comparative analysis in segment 3 fails to identify any true orthologue for dJ525N14.CX.1 and bG421I3.CX.1. In fact, the analysis did not identify any orthologous pairs for any of the human and mouse genes in segment 3.

There are three possible explanations for the lack of synteny observed in segment 3. The first possibility is that the orthologous counterparts of the human genes lie within the gaps present in the mouse sequence. The second possibility is that segment 3 in human and mouse are syntenic to other regions of the mouse and human respectively. However, comparison of the mouse genes with available human genome sequence and human cDNA sequence, and human genes with available mouse genome sequence and mouse cDNA sequence, reveals no other likely candidates for the orthologous genes. A third possibility is that segment 3 in human and mouse are derived from the same region in a common ancestor, but have diverged at a greater rate since the split from the common ancestor than the rest of the region between HPR6.6 and ZNF-Kaiso. This may have occurred if genes in one

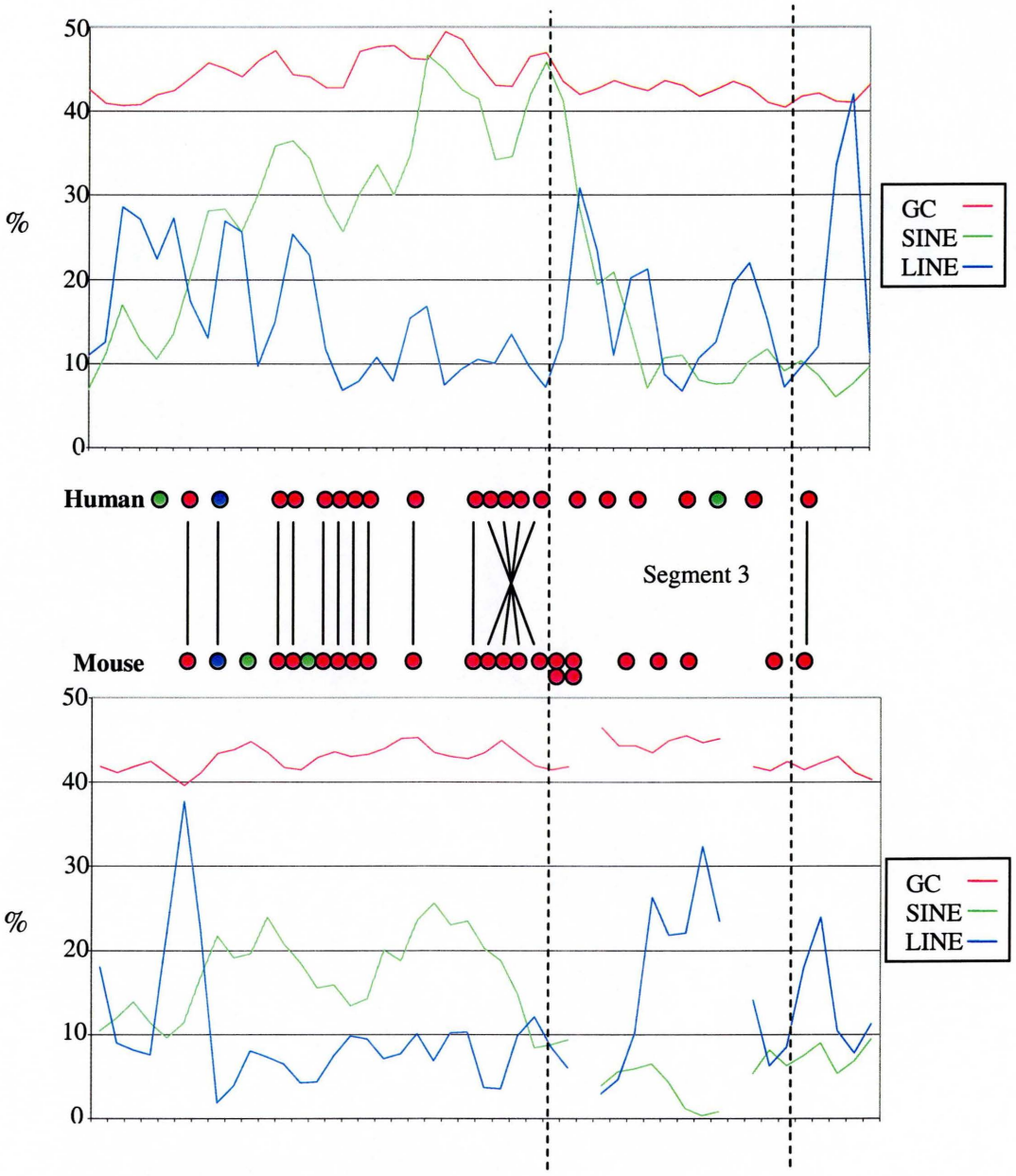
organism had acquired new function. Analysis of segment 3 in other organisms will provide more data to further the understanding of the evolution of the region to support or reject this third explanation.

5.4 Comparison of the genome landscape in human and mouse

Finished sequence in both human and mouse was analysed for GC content and the content of SINEs and LINEs. A series of 50 kb sequence segments overlapping by 25 kb were generated and analysed for repeat content and GC content using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) and the values plotted. The results for the regions in human and mouse were aligned based on the relative positions of the previously identified orthologous genes (see Section 5.3).

Comparison of the genome landscapes in human and mouse shows that there is a correlation between both GC content, and LINE and SINE (see Figure 5.12). The GC content remains above 40% in both human and mouse. Regions of relatively high SINE density were seen to correspond to gene rich regions. In general, the mouse sequence seems to have a lower repeat content, but this may be due to the reduced amount of information currently available for mouse repeats, so that some may remain unidentified. The SINE content both in human and mouse decreases in segment 3, the region where there is no apparent synteny between human and mouse. This correlates with a local change from a gene rich to a gene poor region in both species.

Figure 5.12: *(see over) Comparison of the genome landscape in human and mouse. A 50 kb window, moving in 25 kb increments was analysed for GC content (red line), SINE (green line) and LINE content (blue line), figures are given as a percentage. A break in the line represents a gap in the sequence. The gene content is also shown. Genes shown as red circles, predicted genes shown as blue circles and pseudogenes shown as green circles, orthologous genes are linked with a thin black line. The repeat content is generally lower for the mouse sequence which may reflect the level of understanding of repeat sequences in the two organisms. There appears to be a high SINE content in the region containing the majority of the orthologous pairs of genes, and a lower SINE content in the region containing no obvious orthologous pairs of genes (segment 3 indicated by a vertical dotted black lines).*


50
40
30
20
10
0
%
GC
SINE
LINE
Human
Segment 3
Mouse
50
40
30
20
10
0
%
GC
SINE
LINE

5.5 Analysis of conserved sequences

5.5.1 Evaluating the methods for sequence comparison

One of the major challenges of comparative sequence analysis is to identify the sequences conserved between the two species of interest in any given region, and to elucidate any potential function they might have. As discussed in Section 1.3, regions of DNA that have maintained the same function in different species, since the divergence from a common ancestor are more likely to be conserved at the sequence level than regions that have no function, or have evolved different functions. Therefore, the protein coding regions of the same gene that has maintained the same function and is regulated in the same way is likely to show conservation at the sequence level. Identifying these functionally conserved sequences using sequence comparison is difficult. The region of similarity may be small, given that introns disrupt the similar blocks of sequence, and in the case of regulatory elements, only a few nucleotides may be essential as they bind the regulatory protein.

There have been reports of comparative sequence analysis being used to identity conserved regulatory elements. For example, a region on human 5q31 containing five interleukins and potentially 18 other proteins was compared to draft mouse sequence. Ninety conserved sequences were identified, of which the longest one was shown experimentally to reduce the expression of three of the interleukin genes (Hardison, R. C., 2000). However, it is possible that for regulatory binding sites, the primary sequence composition may not be critical. Instead, other features such as DNA secondary structure may determine the binding activity of the site in some cases, and

these would not be revealed by straight forward sequence comparison (Rubin, E. M., 2001).

Two main methods for aligning two sequences are commonly used, local alignment and global alignment. In a local alignment, regions from one sequence are compared to regions from the other sequence. BLAST is a commonly used local alignment tool and was developed in 1990 (Altschul, S. F., *et al.*, 1990). One of the limitations of the original BLAST algorithm was that for speed, it attempted to extend alignments without introducing gaps in one sequence. More recently, gapped BLAST has been introduced that is able to introduce gaps in the alignment (Altschul, S. F., *et al.*, 1997). In a global alignment, the two sequences are aligned end to end. An example of a global alignment tool is GLASS (Global Alignment SyStem) (Batzoglou, S., *et al.*, 2000). GLASS identifies the longest regions that match exactly, and uses these as a framework to compare the less identical regions in between. Therefore it still has the problem of dealing with gaps and is likely to report separate matches in highly gapped regions, i.e. those region that are least identical.

In this section, three tools are compared for their ability to identify conserved sequences between HPR6.6 and ZNF-Kaiso in human, and Mapr and ZNF-Kaiso in mouse. The region contains a total of 75 known conserved sequences made up of the coding exons of fifteen orthologous genes. Identification of conserved sequences was carried out using PIPMAKER, VISTA and BLAST. PIPMAKER generates local alignments across the region using a version of gapped-BLAST known as BLASTZ (Schwartz, S., *et al.*, 2000). These alignments are then visualised on a percentage identity plot (pip), which shows both the position in one sequence and the degree of

similarity for each aligning segment between two sequences (see Figure 5.13a). VISTA (VISualisation Tool for Alignment), visualises conserved sequences identified using GLASS, and calculating the percent of identical nucleotides within a 100 bp window moved in 25 bp increments across the alignment (Dubchak, I., *et al.*, 2000) (see Figure 5.13b). BLAST allows for rapid sequence comparison by identifying regions of local sequence similarity and for the purpose of this comparison, the alignments were viewed using ACeDB (see Figure 5.13c). Only conserved sequences of greater than 75% identity were scored. The position of the conserved sequences identified by each method is shown in Figure 5.14 and the overall results are compared in Table 5.2.

(a)

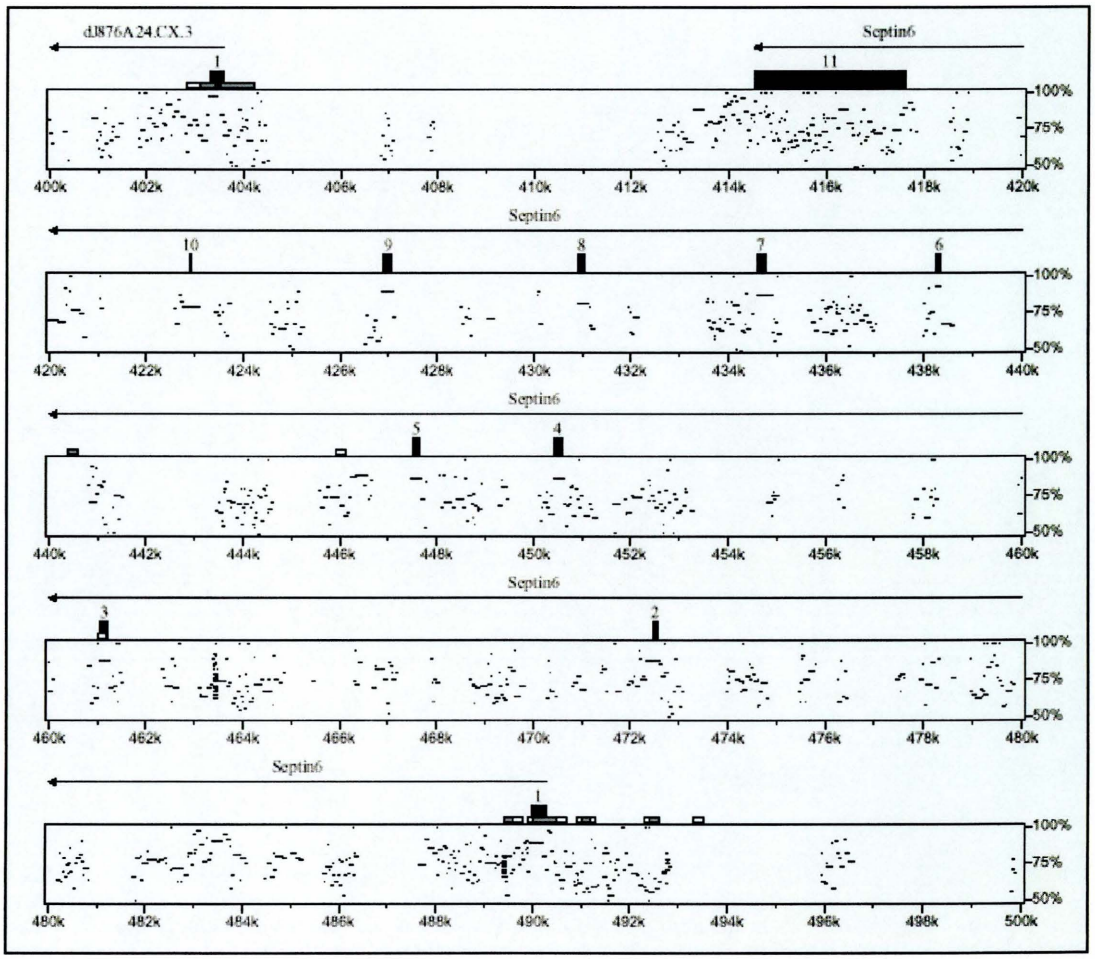


Figure 5.13: *Examples of comparative sequence analysis tools. (a) PIPMAKER. Conserved sequences are identified using BLASTZ and viewed on a percentage identity plot (PIP). Conserved sequences are indicated as black lines positioned relative to their position in the human sequence (horizontal axes) and their percentage identities are shown on the vertical axes. The position of each exon within a gene and the direction of transcription is also given. The example shows a PIP for the region between the 5' end of dJ876A24.CX.3 and the 5' end of Septin6.*

(b)

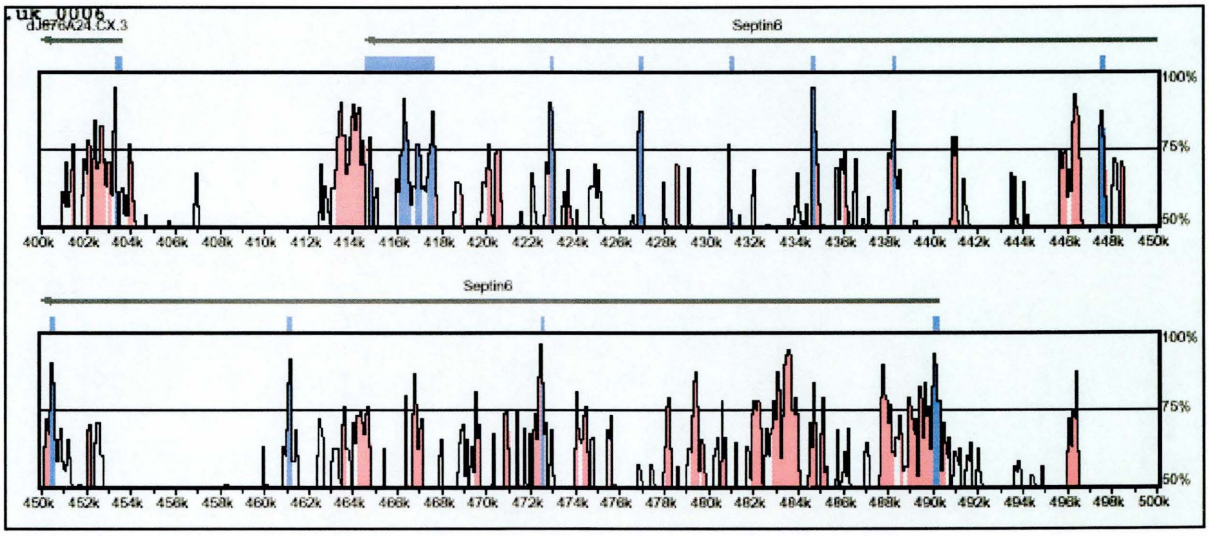


Figure 5.13 cont.: *(b) VISTA. Pairwise sequence alignments are visualised as peaks of similarity. Alignments are generated using GLASS, and conserved sequences identified as a percentage of identical nucleotides within a 100 bp window, moved at 25bp increments across the global alignment of the two sequences. Conserved sequences are indicated as peaks of similarity positioned relative to their position in the human sequence (horizontal axes). Those lying within genes are shown in blue, those lying in introns or intergenic regions are shown in red. The position of each exon within a gene and the direction of transcription are also given. The example shows a VISTA plot for the region between the 5' end of dJ876A24.CX.3 and the 5' end of Septin6.*

(c)

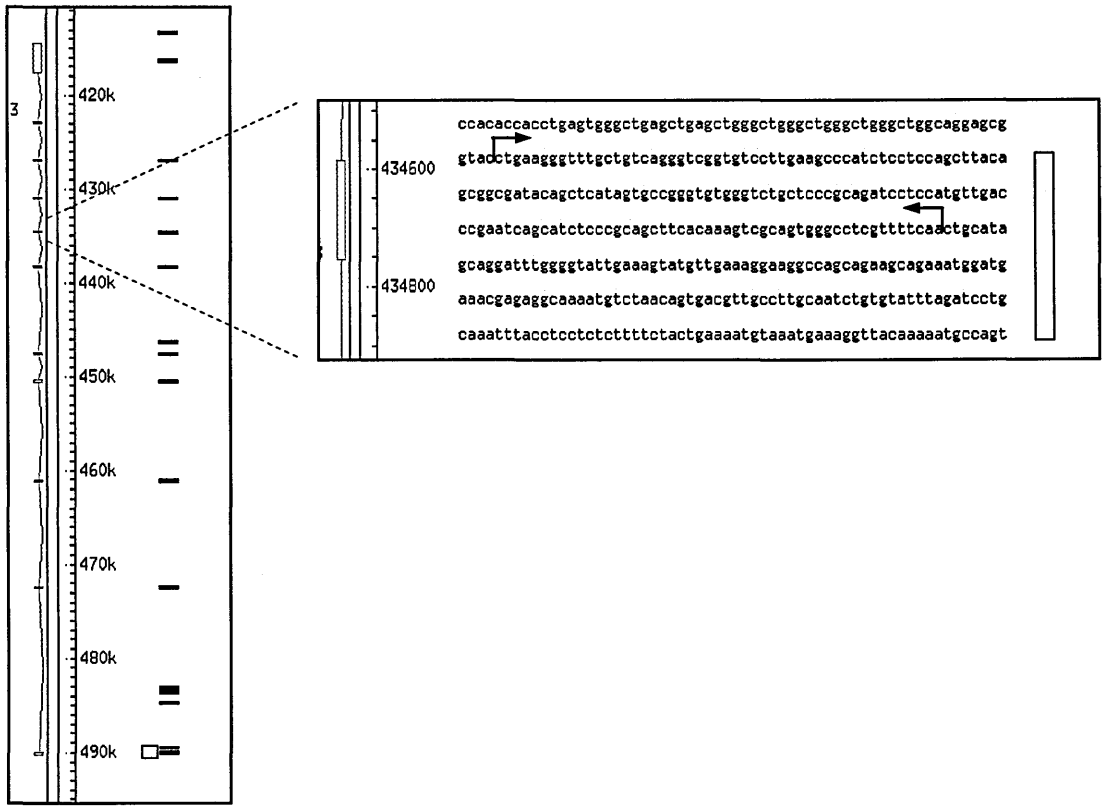


Figure 5.13 cont.: *(c) BLAST and ACeDB. Conserved sequences are identified by BLAST and viewed in ACeDB. A section of the region in human containing the Septin6 (red boxes linked by red lines) is shown and is positioned on the minus strand of the genomic sequence (depicted as a vertical yellow bar). Conserved sequences are shown as black boxes. An example of the sequence of a conserved box is shown (inset). The extent of the alignment is indicated by the highlighted region of sequence, and the position of the seventh exon of the Septin6 gene is indicated by black arrows.*

Table 5.2: *Results of conserved sequence analysis*

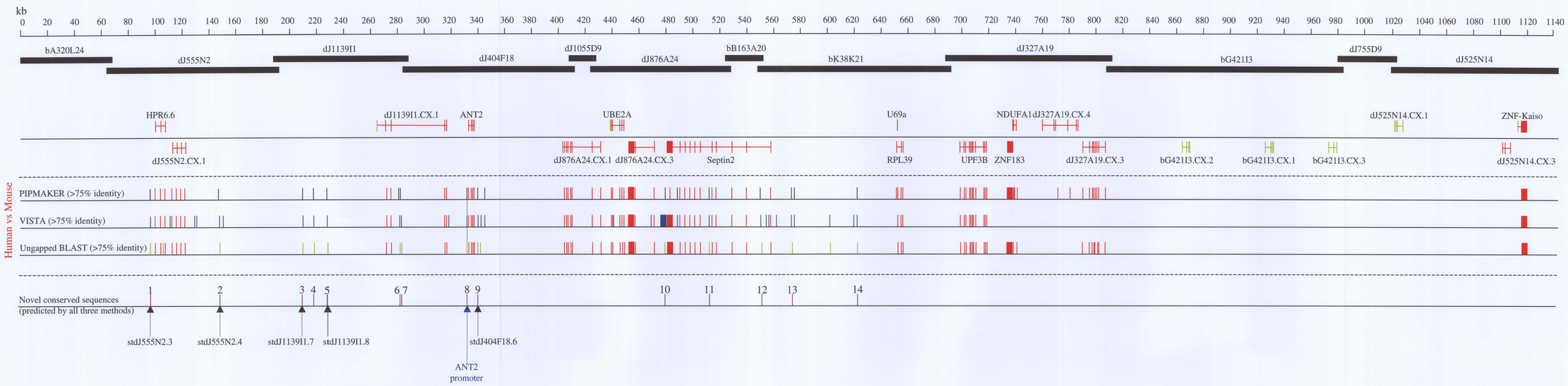
Method	Conserved sequence	Other sequences	Total sequences	Sensitivity	Minimum Specificity
PIPMAKER	71	16	87	0.95	0.82
VISTA	57	30	87	0.76	0.66
BLAST	69	17	86	0.92	0.80
Total	75	30	102	0.96	0.70

Sensitivity is defined as the number of known conserved sequences each method identified as a fraction of the total number of known conserved sequences in the region. Specificity is defined as being the number of known conserved sequences each method identified as a fraction of the total number of conserved sequences predicted by that method. This is given as "minimum" specificity to allow for instances where previously unknown conserved sequences that have been predicted are real.

The results show that PIPMAKER identified 71 of the 75 known conserved sequences (95%), BLAST 69 (91%) and VISTA 57 (76%). VISTA did not identify any of the exons within the four genes that are inverted when compared in human and mouse (see segment 2, Figure 5.9). One of the disadvantages of the global alignment method arises because the sequences are aligned end to end, and the conserved sequences within the genomic rearrangement are not identified. When only segment two was compared between human and mouse using VISTA, a further 14 conserved sequences were identified. These coincided with the same exons predicted by PIPMAKER in the genes ZNF183, NDUFA1, dJ327A19.CX.4 and dJ327A19.CX.4 (data not shown). Using the tools in combination identified only one extra conserved sequence not identified by PIPMAKER alone. Combining the results reduced the specificity to 0.71. These results show that PIPMAKER was the best

method tested for identifying the known conserved sequences between HPR6.6 and ZNF-Kaiso in human and Mapr and ZNF-kaiso in mouse.

Figure 5.14: *Identification of conserved sequences. The extent of the region in human between HPR6.6 and ZNF-Kaiso is shown. A scale in kilobases is given, and the extent of the sequencing of each human clones is shown as black bars. The position of the human genes identified in the region is shown. Each exon is represented by a either a vertical red line or box (red = conserved, as identified by the analysis in Section 5.3, green = not conserved). Genes above the thin black line are transcribed on the plus strand, and those below the line are transcribed on the minus strand. The conserved sequences predicted by three methods are shown below the dotted line. Conserved sequences coinciding with exons are indicated by red lines, those in other regions are shown as black lines for PIPMAKER, blue lines for VISTA, and green lines for BLAST. The position of each STS, designed to the conserved sequences for expression analysis, is indicated. The position of the conserved sequences coinciding with the ANT2 promoter is also shown.*


kb
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500 520 540 560 580 600 620 640 660 680 700 720 740 760 780 800 820 840 860 880 900 920 940 960 980 1000 1020 1040 1060 1080 1100 1120 1140
bA320L24
dJ555N2
dJ1139I1
dJ404F18
dJ1055D9
dJ876A24
bB163A20
bK38K21
dJ327A19
bG421I3
dJ755D9
dJ525N14
HPR6.6
dJ1139I1.CX.1
ANT2
UBE2A
U69a
NDUFA1
dJ327A19.CX.4
dJ525N14.CX.1
ZNF-Kaiso
dJ555N2.CX.1
dJ876A24.CX.1
dJ876A24.CX.3
Septin2
RPL39
UPF3B
ZNF183
dJ327A19.CX.3
bG421I3.CX.2
bG421I3.CX.1
bG421I3.CX.3
dJ525N14.CX.3
Human vs Mouse
PIPMAKER (>75% identity)
VISTA (>75% identity)
Ungapped BLAST (>75% identity)
Novel conserved sequences
(predicted by all three methods)
1 2 3 4 5 6 7 8 9 10 11 12 13 14
stdJ555N2.3
stdJ555N2.4
stdJ1139I1.7
stdJ1139I1.8
stdJ404F18.6
ANT2
promoter

5.5.2 Potential function for novel conserved sequences

One of the difficulties in calculating the specificity for each of the methods is that novel conserved sequences could be real functionally conserved sequences, in which case the initial specificity value would be an underestimate. Twenty-nine novel conserved sequences were identified by at least one method, and fourteen of those were identified by all three methods (see Figure 5.14). One possible function for these conserved sequences is that they represent novel transcribed regions. In an attempt to evaluate this, STSs were designed within five of the conserved sequence regions and screened against the available cDNA resources (marked with a black arrow on Figure 5.14). One STS, stdJ404F18.6, was designed within conserved sequence that was predicted to be within an exon by the exon prediction program HEXON. There was no other evidence to suggest that the other conserved sequences were coding. All STSs were designed within a single open reading frame. For all five of the STSs tested, no expression was observed in the cDNA libraries available (using standard PCR conditions described in Section 2.15.3). Products were seen in some lanes when the cycle number was increased to 40, but subsequent vectorette PCR and sequencing revealed that the sequence did not match the original sequence the STS was designed to. This suggested that the amplified products were generated from the non-specific binding of the primers to cDNA sequences (data not shown).

These results suggest that either these five previously unidentified conserved sequences are not transcribed into mRNA or that they are not represented in the cDNA libraries tested. This evidence is consistent with two other reports that say that only a limited number of novel genes will be identified by comparing human and

mouse sequence, and more commonly, the mouse sequence will provide further definition of genes previously predicted by other methods (Dehal, P., *et al.*, 2001, Deloukas, P., *et al.*, 2001).

A second possibility for the presence of conserved sequences outside genes is that they represent conserved regulatory elements. The regulatory elements for the ANT2 gene have been determined experimentally (Luciakova, K., *et al.*, 2000) (see Figure 5.15). One of the fourteen conserved sequences identified by all three methods (indicated by a blue arrow on Figure 5.14) was observed in the region containing the ANT2 core promoter and one SP1 suppressor element, but no conservation was seen around the two SP1 activating elements. The SP1 supressor element is juxtaposed to the transcription start site, and so the observation of conservation may coincide with the transcription start site and not the SP1 supressor element.

In an attempt to determine whether any of the other conserved sequences could contain regulatory regions, analysis was undertaken using the TRANSFAC database (http://transfac.gbf.de/TRANSFAC/index.html) (Wingender, E., *et al.*, 2000). TRANSFAC is a database of eukaryotic cis-acting regulatory DNA elements and trans-acting factors. The TRANSFAC database contains DNA profiles for each binding site, which can then be used to search stretches of sequence. Each of the 14 conserved sequences predicted by all three methods was searched against the TRANSFAC database using the web site MOTIF (http://motif.genome.ad.jp).

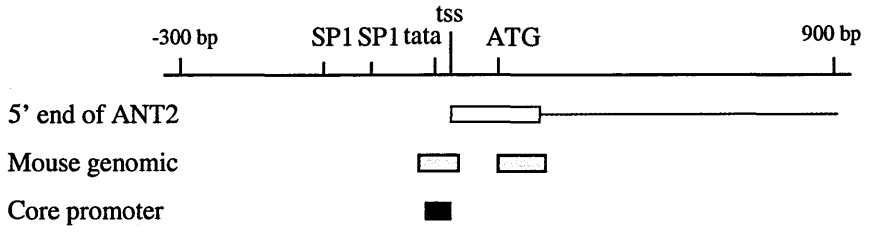


Figure 5.15: *Analysis of the promoter region of the ANT2 gene. The 5' end exon of the ANT2 gene (shown as red box) is shown along with the position of the regulatory elements (tata box (tata), transcription start site (tss) and SP1 activating elements (SP1). The conserved sequences between human and mouse (green boxes) shows conservation within the core promoter region (blue box) and in the coding region (after the ATG), but not in the 5' UTR (defined as lying between the tss and the ATG). No conservation was detected by BLAST around the SP1 activating elements.*

Multiple matches were found in the TRANSFAC database for all of the fourteen conserved sequences. The average size of the conserved sequences is 250 bp and the DNA profiles are often very short sequences (e.g. GATA-2 binding site profile = GATR – where R is purine) and so the probability of finding a match by chance is 1 every 128 bp. However, when the region containing the ANT2 promoter was searched, the correct positions for one SP1 site and the TATA box were identified. Further experimentation will be required to determine whether any of the other thirteen conserved sequences contain real or falsely predicted regulatory elements (see Section 5.7).

5.6 Evaluation of whole genome shotgun (WGS)

The human sequence is nearing completion and the emphasis for large scale sequencing has shifted to generating sequence from genomes of other organisms. In order to produce sequence representing as much of the mouse genome as possible, an initial whole genome shotgun (WGS) has been carried out (data being generated by the Mouse Sequencing Consortium) and made available via trace repositories and BLAST databases on the WWW (http://trace.ensembl.org, http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?, http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html). There are currently almost 30 million reads available representing five genome equivalents. This is calculated by multiplying the number of reads by the average length of a sequence read, in this case 500 bp, and dividing by the genome size, in this case assumed to be $3x10^9$. The initial aim of the consortium was to produce three genome equivalents in WGS

reads, but this figure was recently revised and extended to produce six genome equivalents. In parallel, a clone by clone sequencing effort is also underway, which in combination with the WGS sequencing will provide high quality finished sequence for the mouse genome by 2005.

In an attempt to evaluate to what extent whole genome shotgun from the mouse will aid human sequence annotation, an analysis of unfinished sequence was carried out to assess the contribution of the different levels of shotgun sequence depth (see Figure 5.16). In order to generate the equivalent of different amounts of coverage from WGS, a varying number of sequence reads representing a series of 'coverage equivalents' were used from four mouse BAC clones bM100G16, bM286I5, bM43O20 and bM38B5. Table 5.3 shows the total number of reads available for each clone and the number of reads required for each 'coverage equivalent'.

Table 5.3: *Comparison of read number for various genome equivalents (RD = restriction digest)*

Clone Name	Size by RD (kb)	1/3x	1x	2x	3x	4x	6x
bM100G16	210	141	424	848	1272	1696	2544
bM286I5	230	152	458	916	1374	1832	2748
bM43O20	175	117	352	704	1056	1408	2112
bM38B5	180	118	356	712	1068	1424	2136

The figures were calculated based on an average read length of 500 bp. Analysis of the finished sequence for the four mouse clones showed 73 exons were present from 16 orthologous genes.

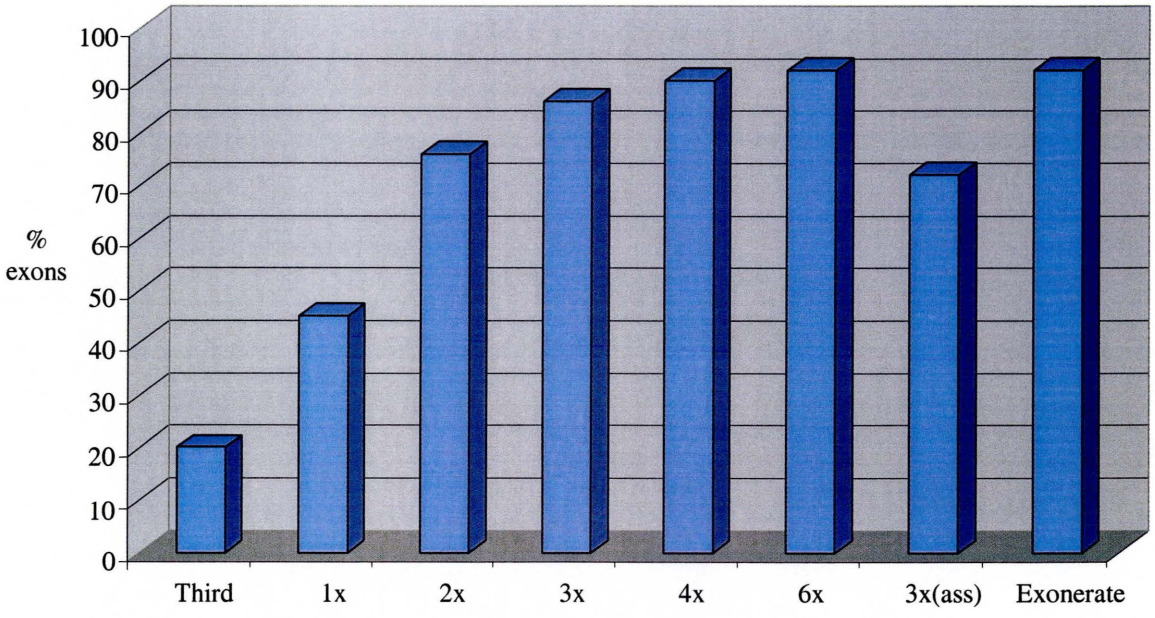


Figure 5.16: *Evaluation of whole genome shotgun. Percentage of matches to human exons in the region (vertical axis) at increasing amounts of coverage of mouse sequence (horizontal axis). 86% of all human exons are present in three genome equivalents of mouse sequence, and 92% are present in six genome equivalents (3x(ass) = 3x assembled). The final bar represents the amount of exons hit by mouse sequence traces currently positioned in the region by EXONERATE.*

The sequence reads from each 'coverage equivalent' were then compared using BLAST to the total number of exons covered by the four clones. The results are shown in Figure 5.16. As the number of reads increases the number of exons present in the mouse sequence increases. This information would suggest that the original target of three genome equivalents of whole genome shotgun data would contain 86% all exons, whereas the revised target of six genome equivalents would contain 92% of all exons. Even though the original reads from the bacterial clones were randomised initially, they were still constrained to lie within a single clone and so cannot be considered precisely comparable. However, analysis of the whole genome shotgun data currently available (approximately five genome equivalents) using EXONERATE, a program that aligns mouse sequence reads to human sequence (courtesy of Michelle Clamp), shows that 92% of the exons are present (see Figure 5.16). A very similar set of exons were identified by both methods, EXONERATE identifying six exons not observed in the sequence generated by the clone based evaluation, which in turn identified six that were not detected by EXONERATE.

5.7 Discussion

Two mouse-specific bacterial clone contigs containing 98 BAC clones covering 1.9 Mb between MAPR and ZNF-kaiso have been generated. The remaining gap has been sized at approximately 50 kb by fibre-fish. In this study, the generation of bacterial clone contigs across the syntenic portions of the mouse genome relied on the identification of a sufficient number of orthologous sequences in mouse cDNA resources. A second approach has been developed using the available mouse BAC end sequence and the human genome sequence (Simon Gregory, personal communications). Using this method the human sequence between HPR6.6 and ZNF-kaiso was compared using BLAST to a database of mouse BAC end sequences from the RPCI-23 and RPCI-24 libraries (generated by TIGR). The analysis identified the same set of BACs as was identified by the hybridisation method described in Section 5.2. By this time, fingerprints were available for all the identified BACs and so the contigs described in this chapter could be constructed without the need for identification of mouse-specific expressed sequences known to be orthologous to the region in human (a similar analysis has been carried out, comparing the whole of human genome to the mouse genome and is described in Section 7.1).

The mouse sequence generated from the minimum set of BAC clones between Mapr and ZNF-kaiso, was shown to contain twenty-three genes and two pseudogenes. Comparison of this region to the syntenic portion between HPR6.6 and ZNF-kaiso in mouse has identified 16 pairs of orthologous genes. The proximal portion of the region, between HPR6.6 and UPF3B in human, appears to be entirely syntenic with

the proximal portion in mouse, between Mapr and Up3b. There is evidence of an inversion of four genes in either human or mouse since the divergence from a common ancestor. The region between bG421I3.CX.2 and dJ525N14.CX.3 in human does not appear to be syntenic with the equivalent region in mouse between bM43O20.CX.8 and bM202F23.CX.3 as no orthologous pairs could be identified.

The region of human sequence studied between HPR6.6 and ZNF-Kaiso contains one gene with a 5' end that could not be confirmed by cDNA sequence (dJ1139I1.CX.1) and one predicted gene for which no cDNA could be detected in the available resources (dJ555N2.CX.1). In both cases, mouse sequence greater than 90% identical across predicted exons was identified (see Figure 5.17). Although the predicted transcribed regions are still to be confirmed with human cDNA sequence, the identification of the mouse orthologue for each gene will provide added confidence to the presence of a real exon or gene. This data will also enable further analyses to be carried out using a wider variety of mouse tissues.

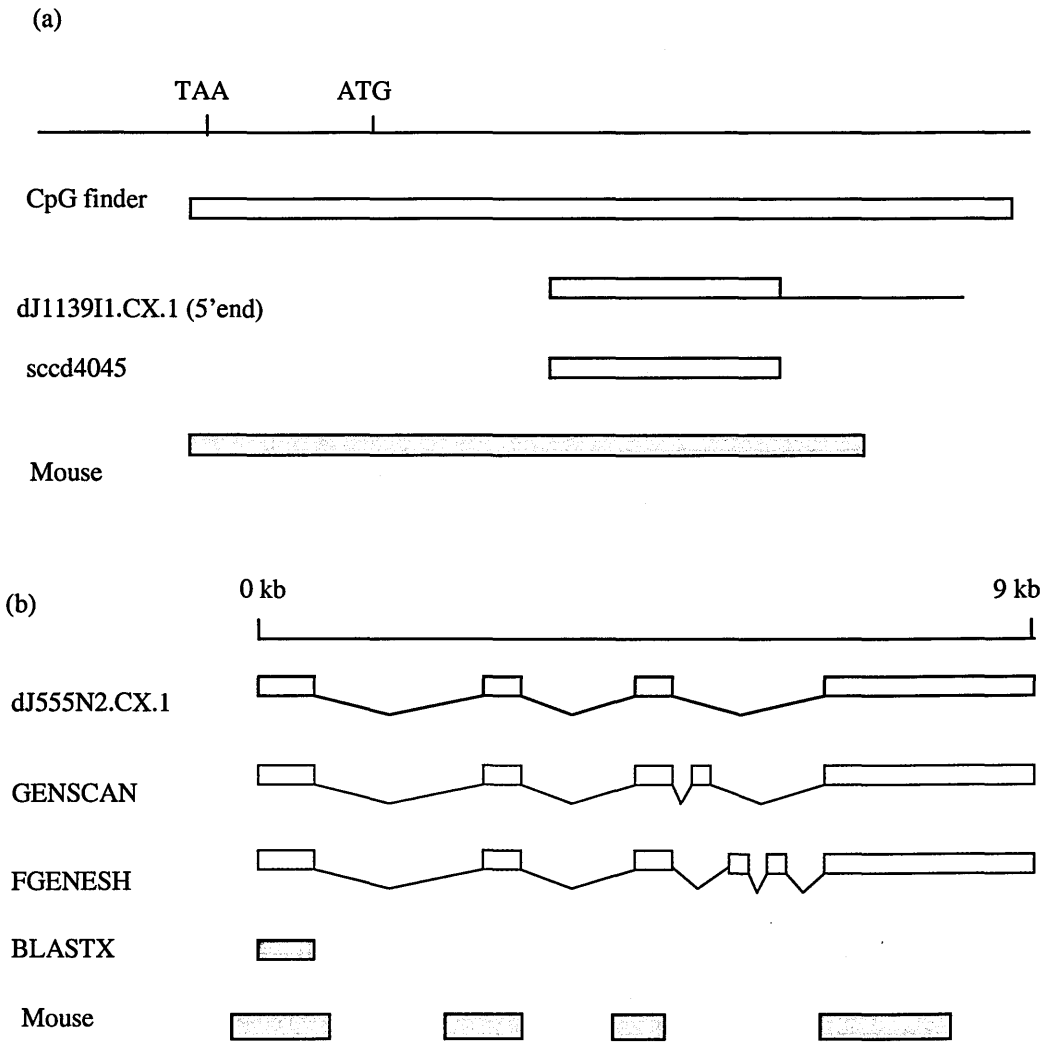


Figure 5.17: *Analysis of predicted and incomplete genes. (a) The 5' end of dJ1139I1.CX.1 (shown as red box). The most likely start site (ATG) is upstream of the confirmed cDNA sequence (pink box). A predicted CpG island is shown as a yellow box. Conserved sequence between human and mouse (green box) extends past the ATG site. (b) The predicted gene dJ555N2.CX.1 (red boxes linked by black lines) and the evidence supporting the prediction (GENSCAN prediction shown as open red boxes linked with red lines, FGENESH prediction shown as open blue boxes linked with blue lines, BLASTX match shown as filled blue box). Conserved sequences between human and mouse (filled green boxes) align with the predicted exons.*

One of the major challenges of comparative genome analysis is the identification and functional analysis of sequences that are conserved between different organisms. Conserved segments could be genes, regulatory elements or other biologically important features such as origins of replication. It is also possible that not all biologically important regions will necessarily show conservation at the primary sequence level, e.g. non-coding RNA genes and regulatory elements. Fourteen conserved sequences of unknown function were identified in the region by three different methods for comparing DNA, PIPMAKER, VISTA and BLAST. Evaluating whether any of these sequences are expressed in a wider variety of cDNA resources in both human and mouse may show that some of these conserved sequences are parts of novel genes.

The conserved sequences may also represent regulatory elements. It has been shown that regulatory elements are conserved in a number of species such as human, mouse and chicken (Gottgens, B., *et al.*, 2000). Experimental analysis is required to determine whether any of these conserved regions function as regulatory elements. For instance, DNA from each conserved region could be cloned into an expression vector in order to test for promoter or enhancer activity. Observing conservation of the regions in other species, such as other mammals or other vertebrates will increase the confidence that these regions are functional.

5.8 Appendix

Table 5.4: *Information on clone names and links shown in Figure 5.8*

Link/Status	Accession	Clone Name
Link_bM100G16	AL450397	RP23-100G16
	AL450399	RP23-286I5
	AL589767	RP23-141L16
	AL451076	RP23-451076
Draft	AL589623	RP23-111C11
Draft	AL590629	RP23-202F23
Draft	AL123456	RP23-322E15
Finished	AL450391	RP23-38B5

Chapter 6

Comparative Sequence Analysis

Between Human and Zebrafish

6.1 Introduction

The identification of human genes and their orthologous counterparts is greatly facilitated by the generation of genomic sequences across the syntenic regions in model organisms (as discussed in previous chapters). As with genomes of other vertebrates, the gene complement of the zebrafish is also expected to show extensive similarity to that of man, thus assisting the annotation of the majority of human genes. The zebrafish genome is approximately 1.7 Gb in size and is divided into twenty-five linkage groups or chromosomes. Recent studies to place zebrafish ESTs onto linkage groups by RH mapping have shown that there is extensive synteny between the human and zebrafish genomes. Pairs of genes in the same region of the human genome are being observed in the same region of the zebrafish genome (Barbazuk, W. B., *et al.*, 2000; Gates, M. A., *et al.*, 1999; Postlethwait, J. H., *et al.*, 1998). However, little is known about how the distances between these pairs of genes differs between human and zebrafish Given that the zebrafish genome is approximately half to two thirds smaller than the human genome, the distance between genes may be smaller in the zebrafish than in the human.

In addition to organisms such as the mouse, the fly and the frog, the zebrafish is one of the organisms of choice in developmental biology, as it is easy to keep, has a short generation time and produces conveniently transparent embryos (Metscher, B. D., *et al.*, 1999). Humans and zebrafish are thought to share a common ancestor, probably a bony fish, which existed approximately 400 million years ago compared to the estimate of 70 million years since the divergence of man and mouse (O'Brien, S. J., *et al.*, 1999). Therefore it is expected that the extent of synteny between human and

zebrafish is less than for human and mouse given the greater time available for gross chromosomal rearrangements in the respective genomes. Current estimates suggest that there are greater than 1000 homology segments between the human and zebrafish genomes (see Figure 6.1, Johnson unpublished), which compares to 200 segments seen between human and mouse (Hudson, T. J., *et al.*, 2001). There are currently 12 separate segments showing homology to the human X chromosome on eight different zebrafish linkage groups (indicated with arrows on Figure 6.1). In this chapter the region between genes HPR6.6 and ZNF-Kaiso in human has been targeted for investigation in the zebrafish, in order to further the understanding of the syntenic relationship between the region of interest in human, mouse and zebrafish, and to identify novel orthologous genes in the zebrafish.

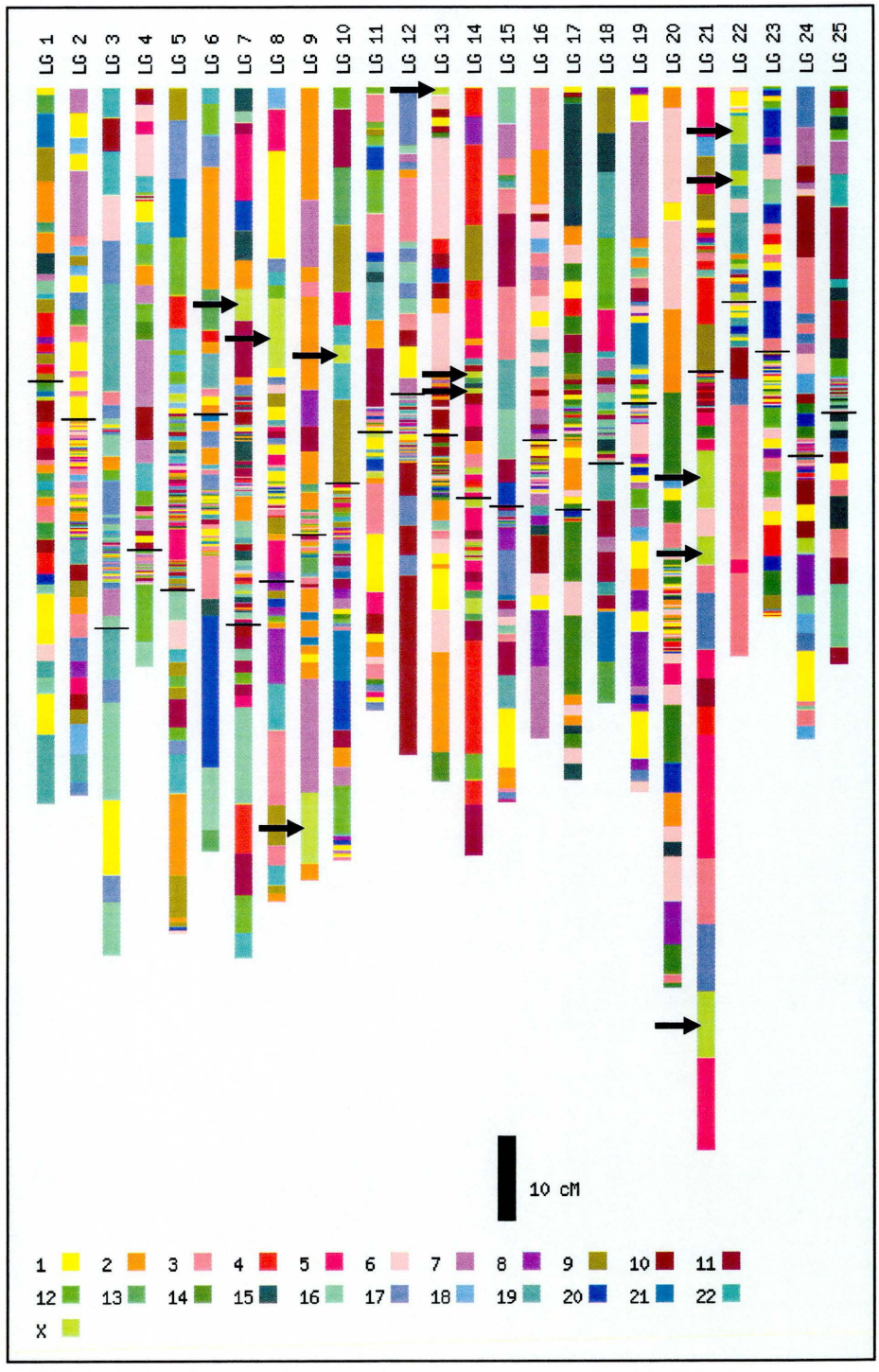


Figure 6.1: *Synteny between human and zebrafish (courtesy of Steve Johnson). The 25 linkage groups (LG) of the zebrafish genome are represented and coloured depending on the positioning of zebrafish-specific ESTs that significantly match human genes. Each colour represents a different human chromosome. The arrows indicate the position of the 11 regions showing synteny to the human X chromosome.*

RESULTS

6.2 Identification of zebrafish genomic clones

In the previous chapter, the strategy for mouse bacterial clone isolation relied upon the knowledge of the sequences of a number of orthologous pairs of genes from which STSs specific to mouse sequences could be designed and used for library screening. At the time this project began, there were no zebrafish sequences known that were orthologous to the human genes in the region between HPR6.6 and ZNF-Kaiso. Therefore a strategy for clone isolation was designed based on using human probes to isolate zebrafish clones by reduced-stringency hybridisation. Fifteen primer pairs were designed within a single exon of sixteen out of of the eighteen genes in the region (as discussed in Section 4.3.4, dJ525N14.CX.1 and bG421I3.CX.1 are 99% identical and a single STS was designed that represented both genes – thus fifteen primer pairs for sixteen genes). No primer pair was designed for two genes, dJ555N2.CX.1 and dJ525N14.CX.3 as these were identified since the clone isolation was carried out. Each STS was labelled and hybridised individually to filters of the zebrafish BAC library (RPCI-71) at 50°C for 16 hours.

A series of washes of increasing stringency (see Section 2.17.3) were carried out. Initial washing was carried out at 50 °C in 6x SSC, 1% Sarkosyl for 2x 30 minutes, and stringency was increased by decreasing the amount of SSC in subsequent wash solutions (4x SSC, 2x SCC, 1x SSC, all with 1% Sarkosyl, and at 50°C for 2x 30 mins). The washing continued until the number of counts present on each filter (measured using a Geiger Counter held to a single filter) dropped below 5 counts per

second. When this point was reached, it was assumed that non-specific binding of probe to the filters had been removed and any probe still bound would potentially represent a sequence-specific positive signal. An X-ray film was exposed to the filters for 36 hours at room temperature. A summary of the results is shown in Table 6.1, columns 1-4. Column 3 shows that for different probes, the filters were washed to different stringencies. For the probe derived from the human UPF3B gene the filters were washed to 4xSSC, whereas for the probes derived from ANT2, UBE2A, RPL39 and NDUFA1, the filters were washed to 1xSSC.

Table 6.1*: Summary of Bacterial Clone Isolation*

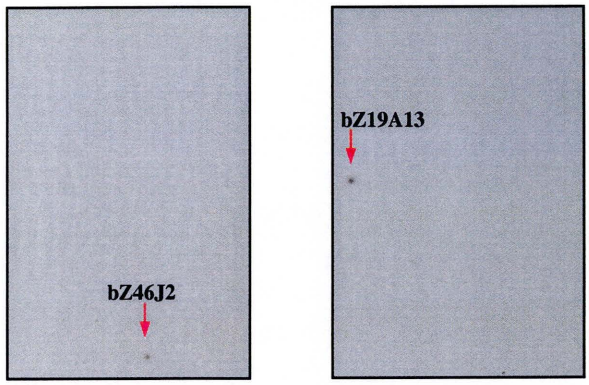
Gene	STS	Wash stringency	No. of clones identified	Clones in ctgs by fingerprinting	No. of ctgs	Sequence clone
HPR6.6	stdJ555N2.2	2xSSC	4*	-		-
dJ555N2.CX.1	-	-	-	-	-	-
dJ1139I1.CX.1	stdJ1139I1.6	2xSSC	4	4	1	bZ21D15
ANT2	stdJ404F18.4	1xSSC	2	2	1	bZ46J2
dJ876A24.CX.1	stdJ404F18.5	2xSSC	5	5	2	bZ80I7 bZ3C13
UBE2A	stdJ876A24.17	1xSSC	3	3	1	bZ46J2
dJ876A24.CX.3	stdJ876A24.16	2xSSC	2	2	0	bZ10G3 bZ20I5
SEPTIN2	stdJ876A24.11	-	-	-	-	-
RPL39	stbK38K21.3	1xSSC	1	1	0	bZ74M9
UPF3B	stdJ327A19.10	4xSSC	3	3	1	bZ79P20
ZNF183	stdJ327A19.12	-	-	-	-	-
NDUFA1	stdJ327A19.13	1xSSC	2	2	1	bZ36D5
dJ327A19.CX.3	stdJ327A19.11	2xSSC	2	2	0	bZ18K17 bZ74M9
bG421I3.CX.2	stbG421I3.4	2xSSC	7	2	1	bZ5O12
dJ525N14.CX.1 bG421I3.CX.1	stbG421I3.5	2xSSC	3	0	0	bZ30I22 bZ71M17 bZ74M9
dJ525N14.CX.3	-	-	-	-	-	-
dJ525N14.CX.4	stdJ525N14.11	2xSSC	1	1	0	bZ39A15

* no clone chosen for sequencing

Thirteen of the fifteen STSs identified a total of 33 positive clones which were assembled into eight contigs by *Hin*d III restriction digest fingerprinting (see Section 2.12.3). A summary of the contigs is given in Table 6.1, columns 5 and 6. On average, one STS identified 2.5 clones which approximately agrees with the estimate that the RPCI-71 library contains three equivalents of the zebrafish genome (RPCI-71 – see http://www.chori.org/bacpac). An example of bacterial clone contig construction using an STS designed to the human ANT2 gene is shown in Figure 6.2. The probe derived from the ANT2 gene identified two clones bZ46J2 and bZ19A13 which when fingerprinted assembled into one contig. Early evidence that two genes, ANT2 and UBE2A, were closely linked in the zebrafish genome came from the fact that the probe derived from UBE2A identified the same clones bZ46J2 and bZ19A13, along with a third clone bZ11A23 which assembled into the same contig by fingerprinting (see Figure 6.2b). This provided supporting evidence that the hybridisation method was identifying sequence-specific signals. A second example of this was seen for two STSs, designed to the genes RPL39 and dJ327A19.CX.3, both of which identified bZ74M9 among other clones (data not shown).

A total of eight clones were identified for sequencing from the eight contigs constructed by fingerprinting (see Table 6.1, column 7). A further six clones were identified for sequencing (shown in red in Table 6.1), from cases in which the STS either identified only one clone, or identified two or three clones that did not show any significant overlap by fingerprinting. For instance, the STS designed to both dJ525N14.CX.1 and bG421I1.CX.1 identified three clones that showed no significant overlap by fingerprinting and all three clones were selected for sequencing.

(a)



(b)

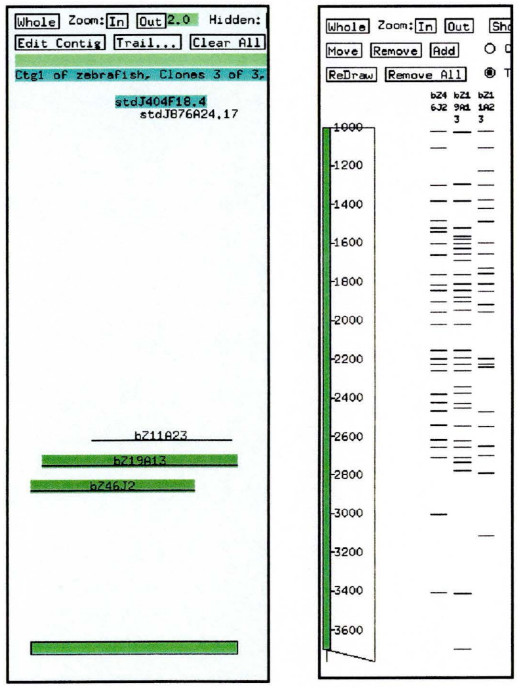


Figure 6.2: *BAC isolation by reduced stringency hybridisation. (a) An example of positive clones when an STS designed within an exon of the human ANT2 gene was hybridised to the zebrafish BAC library. The clones bZ46J2 and bZ19A13 were identified and (b) were assembled into a single contig by Hind III fingerprinting. Both clones along with bZ11A23 (shown un-highlighted) were identified by an STS designed to UBE2A. The fingerprints of all three clones are also given.*

6.3 Evaluation of strategy for the identification of orthologous genes

Analysis of the genomic sequence of fourteen BAC clones by BLAST revealed that only two contained potential orthologous genes. bZ46J2, detected with STSs from UBE2A and ANT2, and bZ74M9, detected with STSs from RPL39 and dJ327A19.CX.3. The remaining 12 BACs did not contain any sequences orthologous to human sequence, and appeared to be false positives identified during the hybridisation procedure. Even though stbG421I3.5 identified bZ74M9, no orthologous sequence was present for this STS. Therefore, it appears that bZ74M9 was identified as a false positive for stbG421I3.5. There was no obvious difference between the signal intensity of real positives versus the false positives. However, analysis of the washing stringency showed that the filters containing the real positive clones were washed to a higher stringency (1xSSC at 50°C) when compared to the false positive clones (greater than or equal to 2xSSC at 50°C).

In order to determine whether increasing the washing stringency could increase the sequence-specificity of detection by hybridisation, multiple filters containing DNA from both the false positive clones and the two real positives were generated. A pooled probe of nine STSs representing ten genes (two STSs that gave real positives and seven STSs that gave false positives) was hybridised to the filters for 16 hours at 50°C. The filters were then washed in steps with increasing stringency, from 4xSSC to 0.1xSSC, at 50°C. After each wash step, one of the filters was removed and stored in 2xSCC at room temperature. The results after exposing the filters to X-ray film for 36 hours at room temperature (the same exposure time as was originally used for the clone isolation) are shown in Figure 6.3.

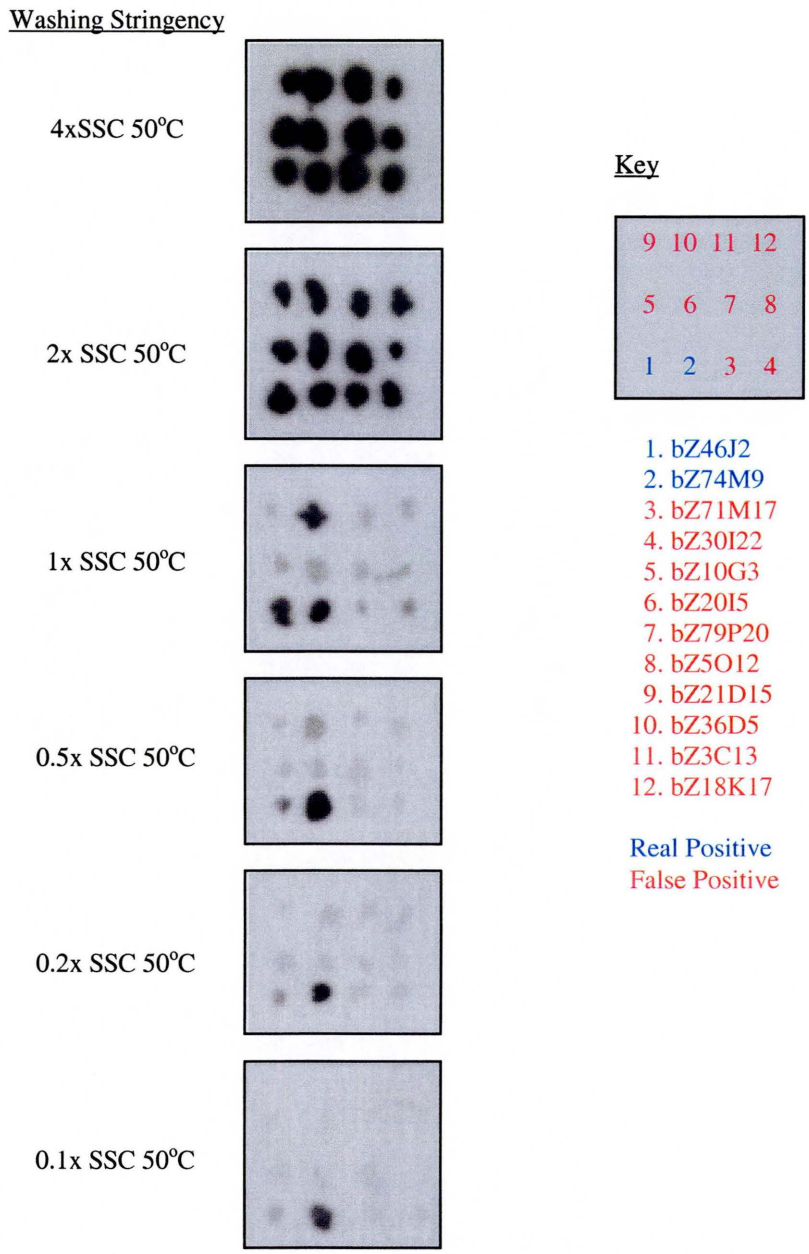
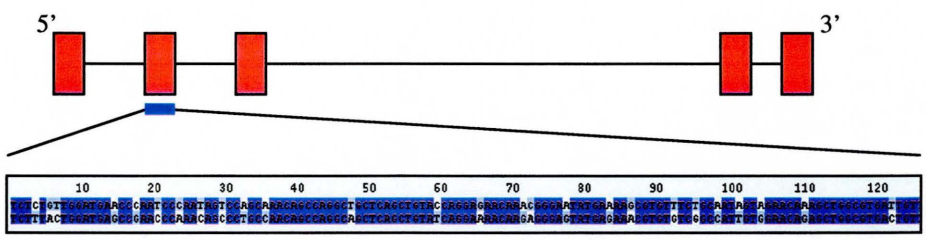


Figure 6.3: *Evaluation of false positives. A pooled probe containing nine STSs was hybridised to six copies of a filter with two true positives (1 and 2) and ten false positives (3-12). The filters were washed with increasing stringency from 4xSSC at 50°C to 0.1xSSC at 50°C. The signal strength remains even for all clones until at 1xSSC only the two true positives, bZ46J2 and bZ74M9, and one false positive, bZ36D5 remain. As the washing stringency continues to increase, the signal of the remaining positives reduces.*

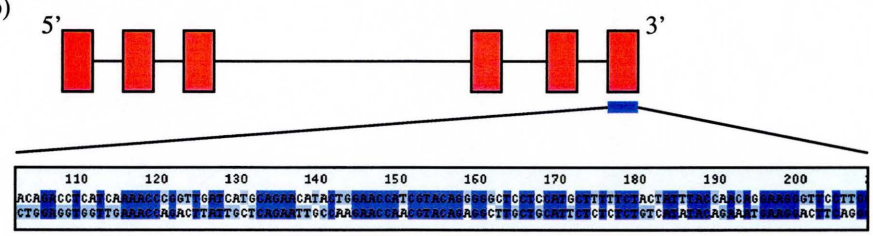
It can be seen that the signals for the real positive clones are still present after washing at 1xSSC, whereas the signal for the false positive clones has all but been removed. One of the clones identified as being a false positive, bZ36D5 was still showing a significant signal even after washing at 1xSSC. Comparison of the sequence of the STS that was designed to the gene NDUFA1, with bZ36D5 by BLAST (Altschul, S. F., *et al.*, 1990) showed that there was a region of 30 bp that was 75% identical between the two sequences. This would be sufficient to account for the apparent sequence-specific signal observed at a wash stringency of 1x SSC (Eric Green, personal communication). Further increasing the stringency of washing by the use of 0.5x SSC shows that the signal from bZ36D5 is removed, but that the signal from the true positive bZ46J2 is also significantly reduced. These results show that washing to a stringency of 1xSSC at 50°C should reduce but not completely eradicate the number of false positive clones in this type of experiment.

Analysis of the two real positive clones (bZ46J2 and bZ74M9) by BLAST showed that they also contained sequences orthologous to four other genes not previously detected by the reduced-stringency hybridisation method. These are dJ1139I1.CX.1, dJ876A24.CX.3, UPF3B and NDUFA1, and the previous negative hybridisation results therefore appear to be false. Analysis of the level of identity between the genomic sequences of the human and the zebrafish in the region of each STS showed that this was higher (above 75%) for those STSs that detected the presence of the orthologous gene by hybridisation, than those that failed to do so (less than 60%) (see Figure 6.4).

(a)



(b)



(c)

Gene	% Identity	Probe Length
UBE2A	79.2	125
ANT2	80.4	256
RPL39	81.5	81
dJ327A19.CX.3	80.8	99
dJ1139I1.CX.1	57.0	105
dJ876A24.CX.3	48.5	101
dJ327A19.CX.1	not matched in sequence	94
NDUFA1	51.4	107

Figure 6.4: *Evaluation of false negatives. (a) Position of the STS (blue bar) designed to human UBE2A (coding exons shown as red bars linked by black lines). The alignment of the human and zebrafish genomic sequence shows the two regions are 79.2% identical over 125 bp. (b) Position of the STS (blue bar) designed to human dJ1139I1.CX.1 (coding exons shown as red bars linked by black lines). The alignment of the human and zebrafish genomic sequence shows the two regions are 57% identical over 105 bp. (c) A table showing the percentage identity between the sequence in human and zebrafish for each STS and the length for the four true positives (names shown in red) and the four false negatives (names shown in blue). The two sequences for each STS are greater than 75% identical for the true positives, and less than 60% identical for the false negatives.*

In summary the technique described here for the identification of zebrafish orthologues of human genes in BAC clones is able to detect the presence of the orthologue in some instances, but the technique is dependent upon the sequence similarity between the probe and the genomic sequence being greater than approximately 75% identical. A reduction in the stringency might reduce the false negative level but would also result in a significant rise in the false positive rate.

6.4. Identification of BAC clones using orthologous zebrafish EST sequence.

At this time, a more detailed radiation hybrid (RH) map of the zebrafish genome, containing the locations of a large number of zebrafish EST sequences was made available (courtesy of Steve Johnson, Washington University, St Louis). Analysis of the region of interest between HPR6.6 and ZNF-Kaiso, showed that two genes appeared to be orthologous to zebrafish EST sequences. dJ876A24.CX.3 matched to EST wz3779 and dJ327A19.CX.3 matched to EST wz8217 (information for each zebrafish EST can be obtained from http://www.genetics.wustl.edu/fish_lab/cgi-bin/display.cgi). These ESTs were positioned at the same point on zebrafish linkage group (LG) fourteen at 14:56 centiRays (cR). A further twenty-one zebrafish ESTs have been mapped to this position in the RH map, but comparison of these sequences with the other human genes in the region by BLAST revealed no other significant matches (data not shown). For the genes HPR6.6 and ZNF183, the potential orthologous zebrafish EST sequences, have been mapped to LG20 and LG7 respectively.

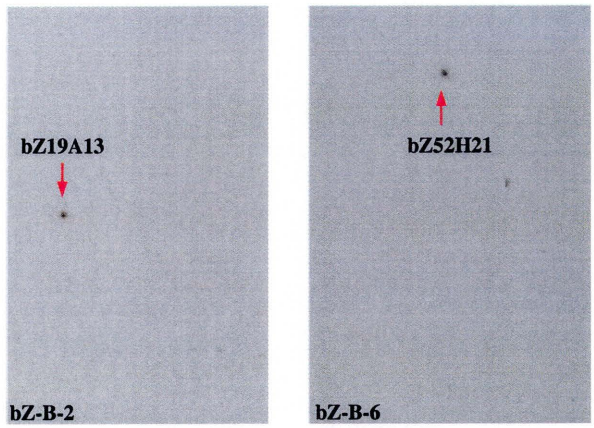
The identification of two zebrafish EST sequences orthologous to the human genes dJ876A24.CX.1 and dJ327A19.CX.3 enabled the use of the method described in chapter 5 for the identification of zebrafish BAC clones (see Figure 6.5). A probe from each zebrafish gene was produced by amplification for the PCR using primer pairs predicted to be within a single exon (based on an alignment of human and zebrafish sequences). The probes were hybridised as a pool to the zebrafish BAC library (RPCI-71). In this case, a total of six BACs were obtained by hybridisation, of which all six were confirmed by PCR. As expected the STSs mapped to the two contigs previously shown in Section 6.3 to contain the zebrafish orthologues of dJ876A24.CX.3 and dJ327A19.CX.3. Three additional clones, bZ52H21, bZ67M17, and bZ62I22 were identified by this method as opposed to the reduced stringency hybridisation method and incorporated into the contigs by fingerprinting. An example of one contig can be seen in Figure 6.5c.

It has been suggested that large regions of the ancestral zebrafish genome may have undergone either total or partial genome duplications (Barbazuk, W. B., *et al.*, 2000). These two zebrafish-specific STSs were used as probes to identify bacterial clones that assembled into two contigs, one contig containing all the clones positive for one STS, the other contig containing all those positive for the other STS. The zebrafish EST sequences to which the two STSs were designed, have been positioned at LG14:56. There is no evidence from these data that the region containing the two zebrafish EST sequences, syntenic to the region in human containing dJ876A24.CX.3 and dJ327A19.CX.3, is present more than once in the zebrafish genome. However, it is still possible that a duplication of the region has taken place but that the sequence of one copy has diverged sufficiently so as not to be detected

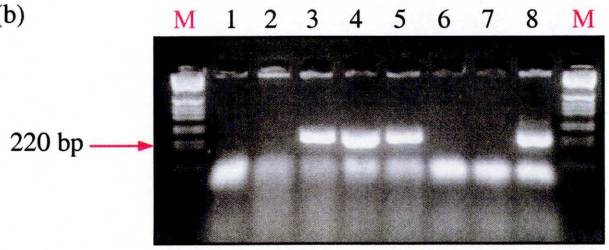
by the method described. Analysis of the complete sequence of the zebrafish genome will enable a more detailed study of the region and an exhaustive search for other homologous regions at a lower stringency for evidence of a possible duplication in the zebrafish that was followed by substantial sequence divergence.

Figure 6.5: *(see over) Identification of BAC clones using an STS designed to the zebrafish EST wz3779 (a) Autoradiograph of two of the six filters (bZ-B-2 and bZ-B-6) after hybridisation of a pool of two STSs, stwz3779.1 and stwz8217.1, and washing to 0.5xSSC at 65°C, showing two positives bZ19A13 and bZ52H21. (b) Colony PCR of the positives from the hybridisation with the stwz3779.1 showing the positive clones bZ46J2, bZ19A13 and bZ52H21. The clones in lanes 1, 2, and 6 were shown to be positive with stwz8217.1. M = Marker. (c) FPC diagram showing the clones identified with stwz3779.1 assembled together by fingerprinting (highlighted in green). One other clone is also present in the contig. bZ11A23 was identified with an STS designed to the human UBE2A (indicated by a red arrow). The fingerprints of the clones are also shown. Bands in the fingerprint for bZ52H21 matching other bands in other lanes are shown in blue, and those matching bands in the other lanes are shown in red.*

(a)

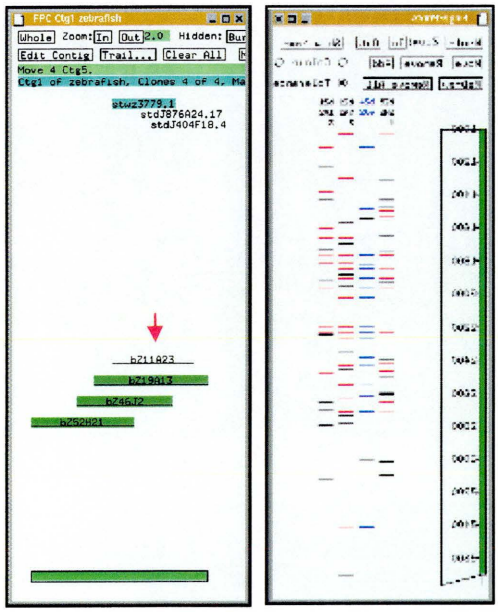


(b)



1. bZ67m17
2. bZ74M9
3. bZ46J2
4. bZ52H21
5. bZ19A13
6. bZ62I22
7. To.1E
8. ZF DNA

(c)



6.5 Sequence Analysis

As discussed in Sections 6.3 and 6.4, bZ46J2 and bZ74M9 were shown by BLAST to contain eight sequences, identified as being orthologous to genes in the region between HPR6.6 and ZNF-Kaiso in human. A more complete analysis of the sequence of bZ46J2 and bZ74M9 has been carried out using a combination of sequence similarity searches and *de novo* gene prediction. The analysis identified a total of twelve predicted genes (see Figure 6.6), but no cDNA-based experimental confirmation of gene structures was carried out due to the lack of available zebrafish cDNA resources at the time.

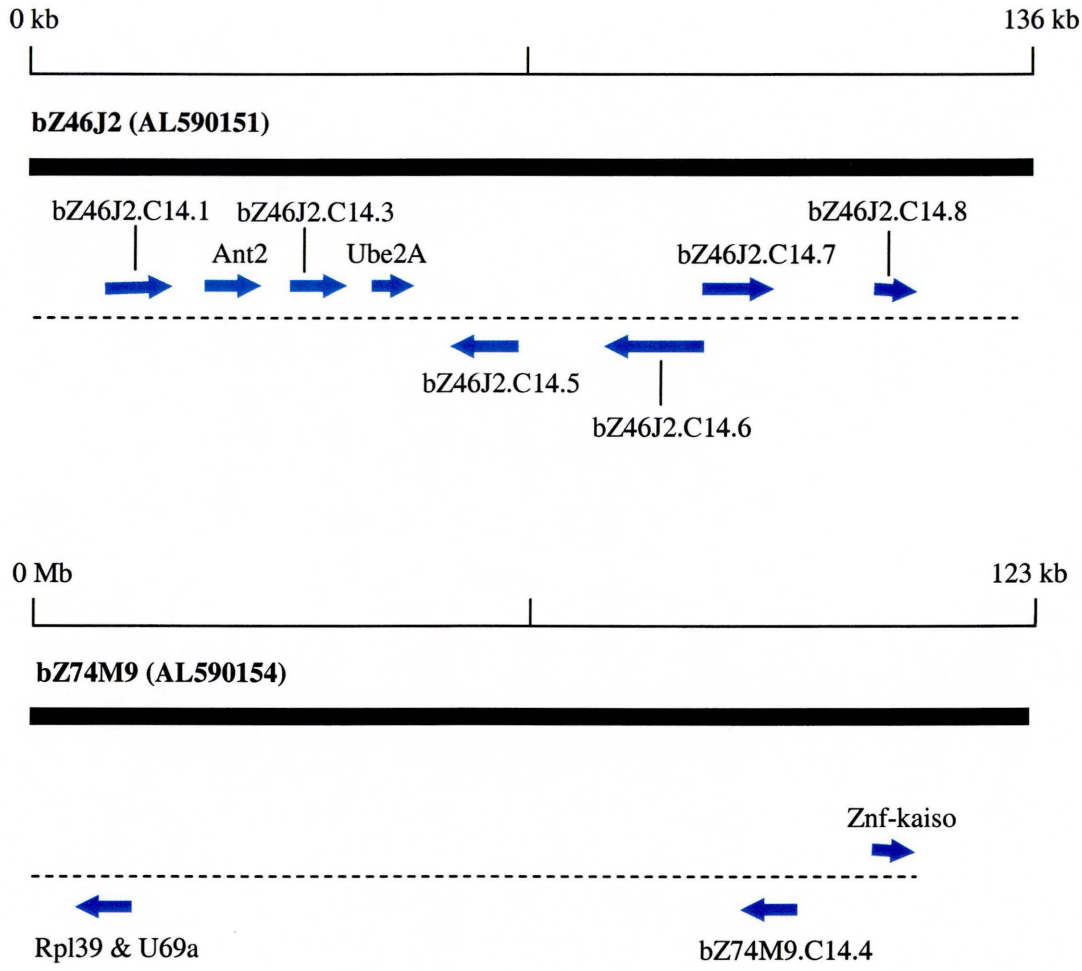


Figure 6.6: *Summary of the gene map constructed in zebrafish. The black bars indicate the finished sequence of the two clones analysed with the accession numbers in brackets. A scale is given in kilobase pairs (kb). Predicted genes are indicated by blue arrows, the direction of each arrow reflects the direction of transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.*

Comparison of the twelve predicted zebrafish genes with the genes in human between HPR6.6 and ZNF-Kaiso, showed that eight were newly identified orthologous pairs, based on their position, similarity at the nucleotide and protein level and similar gene structures (see Figure 6.7 – see also appendix at the end of this chapter for comparison of all eight orthologous genes in human, mouse and zebrafish). For instance, the ANT2 gene in human and mouse has been compared to the newly identified zebrafish orthologue (see Figure 6.8). There is good conservation between the sizes of exons, but less conservation in the sizes of introns. In general, the intron sizes are smaller in mouse and zebrafish along with the distances between genes, which may reflect the differences in the size of the respective genomes (3 Gb in human and mouse, and 1.7 Gb in zebrafish).

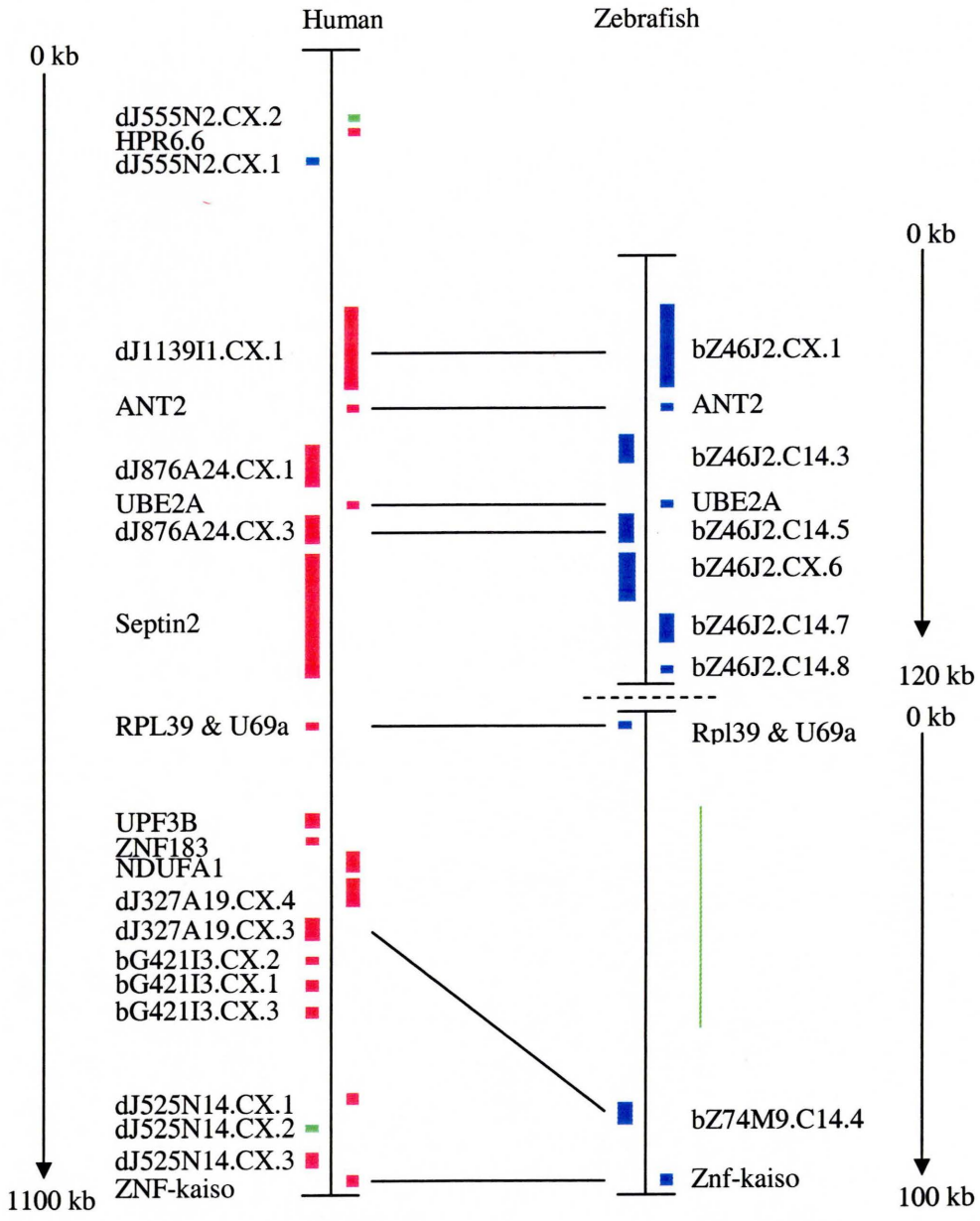
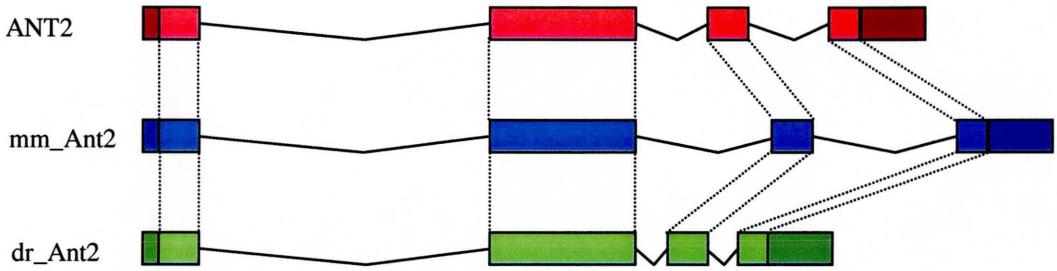


Figure 6.7: *Comparison of the genes identified in zebrafish (on the right) with the genes in the region of interest between HPR6.6 and ZNF-Kaiso in human (on the left). A vertical bar represents the extent of the sequence and genes are shown as bars (red = genes confirmed by cDNA, blue = predicted genes, green = pseudogene). Horizontal black lines link predicted orthologous pairs. A vertical green line indicates the region containing 6 direct repeats. The size of each region is indicated and suggests a tighter clustering of genes in zebrafish than was observed in human.*

(a)



(b)

Gene Name	ANT2	mm_ANT2	dr_ANT2
5'UTR	70	85	87
exon 1	111	111	111
2	487	487	487
3	141	141	141
4	158	158	158
3'UTR	258	263	305
Total Coding	897	897	897
Total cDNA size	1225	1245	984
intron 1	1034	1001	1015
2	225	406	80
3	387	510	84
Total Intron size	1646	1917	1179
Genomic coverage	2871	3162	2163

(c)

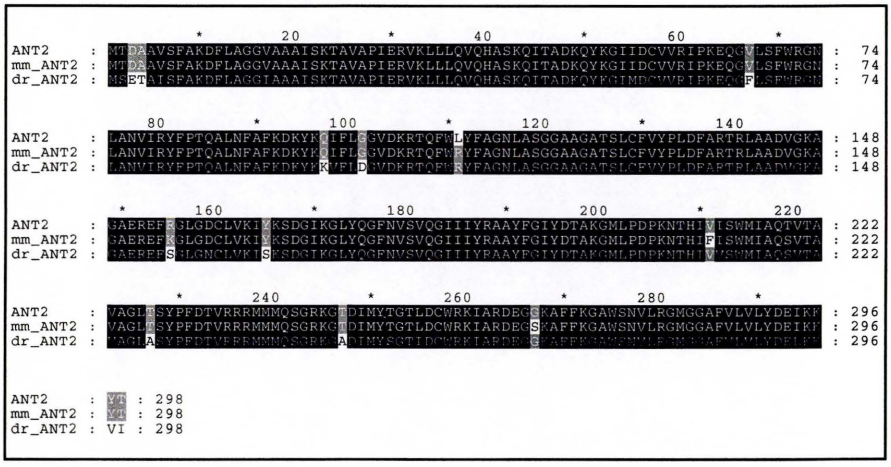
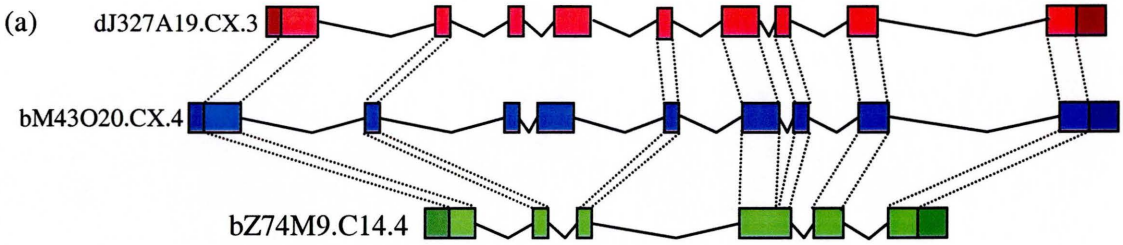


Figure 6.8: *Analysis of orthologues in human, mouse and zebrafish (1) (a) A schematic representation of the ANT2 gene (exons are shown as bars, introns by v-shaped lines) in human (red) mouse (blue) and zebrafish (green). Untranslated regions are shown darker. Dotted lines indicate equivalent exons, based on sequence similarity and exon size. (b) Comparison of exon and intron sizes of the three genes, showing good continuity between sizes of the coding exons. (c) An alignment of the predicted protein sequence of the ANT2 genes in human (top), mouse (middle) and zebrafish (bottom). Amino acids identical in genes from all three species are shaded in black.*

The genes dJ327A19.CX.3 (human), bM43O20.CX.4 (mouse) and bZ74M9.C14.4 (zebrafish) have also been classed as orthologous counterparts of each other (see Figure 6.9). Alignment of the predicted protein sequences of the three genes shows that the encoded human and mouse proteins are 94% identical to each other, and only 60% identical to the zebrafish protein. The 3' ends of the three genes are the part of this homologous set that are most similar to each other. Two exons in the human and mouse genes (exons 6 and 7) are present as a single exon in zebrafish (exon 4). Exons 3 and 4 in mouse do not appear to be present in the zebrafish gene. At this point there is no information regarding the possible function of these proteins, and so no conclusions can be drawn about the effect the amino acid sequence encoded by the extra exons in human and mouse will have on the function of the respective proteins.

Figure 6.9: *(see over) Analysis of orthologues in human, mouse and zebrafish (2) (a) A schematic representation of three genes, dJ327A19.CX.2, bM43O20.CX.4 and bZ74M9.C14.4 (exons are shown as bars, introns by v-shaped lines) in human (red) mouse (blue) and zebrafish (green). Untranslated regions are shown darker. Dotted lines indicate equivalent exons based on sequence similarity and exon size. (b) Comparison of exon and intron sizes of the three genes. (c) An alignment of the predicted protein sequence of the three genes in human (top), mouse (middle) and zebrafish (bottom). The three genes are similar at the 3' end, less similar at the 5' end, and the human and mouse genes encode extra amino acids in the middle of the protein.*

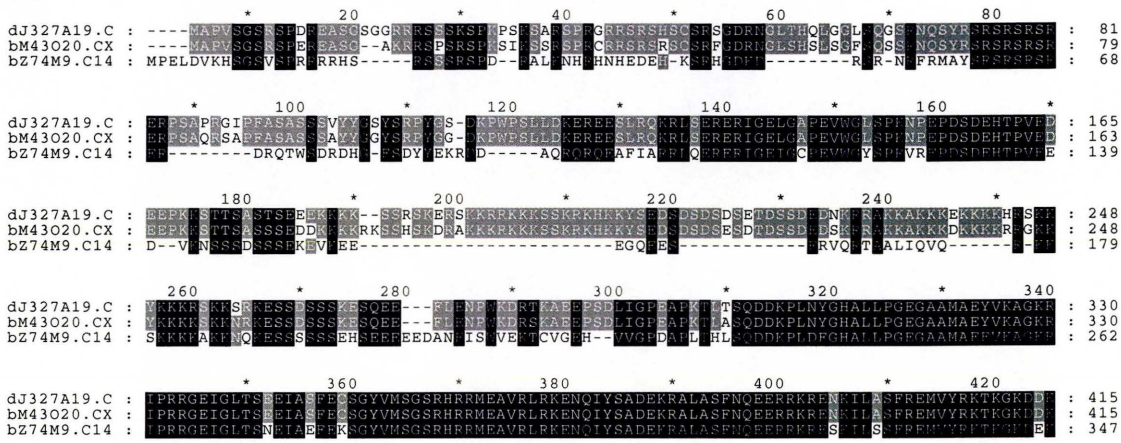
(a)



(b)

Gene Name	dJ327A19.CX.3	bM43O20.CX.4	dr_bZ74M9.C14.4
5'UTR	38	184	262
exon 1	386	380	308
2	81	81	81
3	71	71	64
4	135	141	189
5	64	64	150
6	110	110	175
7	76	76	
8	150	150	
9	175	175	
3'UTR	161	317	276
Total Coding	1248	1248	967
Total cDNA size	1447	1749	1505
intron 1	4409	3602	887
2	2047	5453	541
3	180	169	3319
4	1758	2784	93
5	2277	1897	625
6	74	89	
7	1770	992	
8	4621	4736	
Total Intron size	17136	19722	5465
Genomic coverage	18583	21471	6970
Distance to next gene	67048		

(c)



Four zebrafish genes do not appear to be orthologous to any of the other human genes in the region between HPR6.6 and ZNF-Kaiso, using the criteria described in Section 5.3. It is not unexpected that, due to the evolutionary distance between human and zebrafish, genes located in different regions of the human genome are present in the same region in zebrafish. However three of the four genes, bZ46J2.C14.6, bZ46J2.C14.7 and bZ46J2.C14.8 do match known genes located elsewhere in the human genome. bZ46J2.C14.6 shows similarity to members of the arrestin family of proteins, which are involved in the inactivation of rhodopsin and other heptahelical receptors. A comparison of the predicted protein sequence of bZ46J2.C14.6 with available protein sequences in EMBL using BLAST, shows that the most similar protein in human is β-Arrestin-1 (Sw:P49407) which is 46.15% identical. The predicted protein product of the zebrafish gene bZ46J2.C14.7 is approximately 60% identical to the human Inositol polyphosphate phosphatase-like 2 (INPPL2) protein, and the predicted protein encoded by the gene bZ46J2.C14.8 is approximately 61% identical to a human purinergic receptor (P2RY2). The three human genes are located on human chromosome 11q13 (data taken from ENSEMBL), within a 4 Mb region (82.9 Mb to 86.3 Mb). This analysis would indicate the presence of a syntenic block between human chromosome 11q13 and a region on zebrafish linkage group 14.

The one remaining zebrafish gene, bZ46J2.C14.3 did not match any sequence from any other organism currently available. bZ46J2.C14.3 has four exons and has a predicted mRNA size of 565 bp. The predicted protein is 184 amino acids in length and analysis of the protein in INTERPRO (http://www.ebi.ac.uk/INTERPRO) failed to identify any match to known protein domains.

The zebrafish genes, Rpl39 and bZ74M9.CX.4 are further apart than their predicted human orthologues RPL39 and dJ327A1.CX.3 (see Figure 6.7). Analysis of the genomic sequence in between the two genes in zebrafish reveals a region that may have undergone expansion due to the presence of five zebrafish-specific direct repeats (see Figure 6.10).

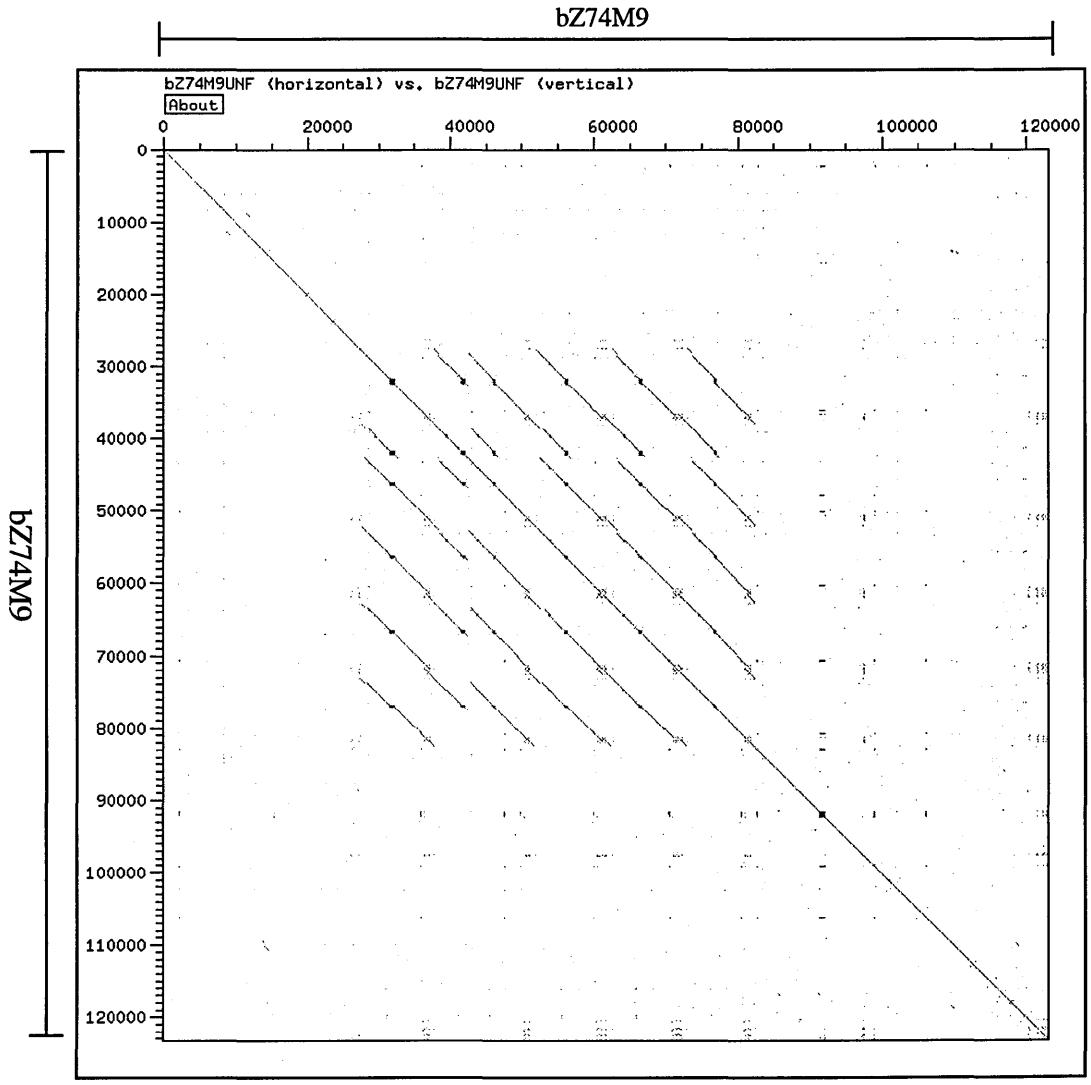


Figure 6.10: *A DOTTER of bZ74M9 against itself showing the presence of five copies of a direct repeat (indicated by red lines).*

6.6 Identification of 20 novel repeat elements in the zebrafish genome

The BAC clones identified in this chapter were among the first zebrafish clones to be sequenced by the Sanger Institute Sequencing teams. In order to further the understanding of the repeat content of the zebrafish genome the genomic sequence was analysed for the presence of repeats. At the time there were nine zebrafish repeat sequences listed in REPBASE (http://www.girinst.org) and these are summarised in Table 6.2.

Table 6.2: *Breakdown of known repeats*

Repeat Name	Repeat Type	Reference
ANGEL	DNA transposon	Izsvak, Z., *et al.*, 1999
BHIKHARI	DNA retroposon	Vogel, A. M., *et al.*, 1999
BHIKHARII	DNA retroposon	Vogel, A. M., *et al.*, 1999
BRSATI	Satellite type I DNA	Ekker, M., *et al.*, 1992
DANA	DNA retroposon	Izsvak, Z., *et al.*, 1996
DRSATII	Satellite type II DNA	Ekker, M., *et al.*, 1992
LINE_DR	LINE-like	direct submission
TDR1	Tc1-like element	Izsvak, Z., *et al.*, 1995
TZF28	DNA transposon	direct submission

In an attempt to identify novel repeat sequences in the zebrafish genome, the draft sequence from the fourteen available BAC clones were compared to each other by BLAST and the clones analysed for regions of sequence that were present three or more times. These regions are candidates for novel repeat sequences and a consensus of each novel repeat region was generated (courtesy of Sarah Hunt). A total of twenty novel repeat sequences have been identified and these are summarised in Table 6.3.

Table 6.3: *Summary of novel repeat sequences in the zebrafish genome*

Repeat Name	Repeat Length	Matches in genome	Sequence contribution (kb)
DR_Rep1	1139	21	23.9
DR_Rep2	578	104	60.1
DR_Rep3	522	108	56.4
DR_Rep4	526	20	10.5
DR_Rep5	735	33	24.3
DR_Rep6	238	29	6.9
DR_Rep7	191	6	1.1
DR_Rep8	110	208	22.9
DR_Rep9	1407	100	140.7
DR_Rep10	670	75	50.3
DR_Rep11	198	150	29.7
DR_Rep12	485	16	7.7
DR_Rep13	391	6	2.3
DR_Rep14	593	3	1.7
DR_Rep15	908	75	68.1
DR_Rep16	375	210	78.7
DR_Rep17	110	67	7.3
DR_Rep18	1226	300	367.8
DR_Rep19	555	125	69.4
DR_Rep20	1117	222.5	2485.3
Total	-	-	3515.1

In order to get an estimate for the number of copies of each repeat in the zebrafish genome, each repeat was compared to the available zebrafish whole genome shotgun sequence using the Trace Archive available at http://www.ncbi.nlm.nih.gov/Traces/trace.cgi. At the time the analysis was carried out there were 4.2 million traces deposited which represents approximately 1.2 genome equivalents (assuming an average read length of 500 bp, and a genome size of 1.7 Gb). The number of copies for each repeat was calculated as being the number of different matches across the entire length of the repeat, divided by 1.2 (the genome equivalents available). The results are shown in Table 6.3, column 3.

Using this information, it is possible to estimate the amount of DNA sequence these novel repeats contribute to the zebrafish genome (see Table 6.3, column 4). Based on the length of each novel repeat and the number of copies in the genome, the data would suggest that these novel repeats contribute approximately 0.2% to the zebrafish genome size. A similar analysis was carried out using the previously known repeat sequences and this showed that they contributed 0.06% of the zebrafish genome.

The zebrafish genome is a half to two thirds smaller than the human genome which may be accounted for in part by a lower repeat content. However, the zebrafish genome is more than four times larger than *Fugu rubripes* genome (1.7 Gb compared to 0.4 Gb for fugu) and therefore may be expected to contain a greater number of repeat sequences. It is known that the genome of *Fugu rubripes* contains very few repeat sequences which is thought to account in part for the reduced genome size, along with reduced intron sizes (Elgar, G., *et al.*, 1999). The number of repeats identified in this study and the estimates for the percentage of the zebrafish genome that is made up of repeat sequences is likely to be an underestimate because of the limitations of the analysis described here. Fourteen clones is a very small number to identify novel repeats in. These may be biased to regions of high or low repeat content. Also, the novel repeat sequences have been compared to short sequences of approximately 500 bp available as individual sequence reads from the whole genome shotgun of the zebrafish genome, which may generate errors in the analysis.

Further investigation of these repeats on large stretches of finished zebrafish sequence from different regions will improve the analysis and the continuing efforts of sequence generation will allow this. Also, it is possible that the repeats contained within the zebrafish genome are highly diverged and re-iterative BLAST analyses with all the repeat sequences may identify more copies of each novel repeat sequence. This will allow for a more detailed study of the repeat content and enable comparisons to be made with other fish genomes, such as fugu whose repeat content is reported to be very low, and mammals such as human and mouse.

6.7 Multiple sequence analysis

Genomic sequence in mouse and zebrafish has been generated that appears to be syntenic to parts of the region between HPR6.6 and ZNF-Kaiso in human. A total of twelve genes have been identified in zebrafish, of which eight appear to be orthologous in three different species, human, mouse and zebrafish (see Figure 6.11). Pairwise analysis of the sequence in human and mouse (see chapter 5) revealed a total of twenty-nine novel conserved sequences predicted by at least one of PIPMAKER, VISTA or ungapped BLAST and fourteen of those were predicted by all three. The additional information provided by the zebrafish sequence generated in this chapter allows for a further evaluation of the conserved sequences in the region between HPR6.6 and ZNF-Kaiso. Comparisons were carried out between human and zebrafish using the same three methods described in the previous chapter. Conserved sequences can only be identified for part of the region in human between HPR6.6 and ZNF-Kaiso given the lack of sequence covering the entire syntenic region in

zebrafish. Given the increased evolutionary distance between human and zebrafish, the threshold for sequence similarity was reduced from 75% (used for human-mouse comparisons) to 50%. The results are shown in Figure 6.12 and summarised in Table 6.4.

Figure 6.11: *(see over) Comparison of genes identified in human (middle), mouse (right) and zebrafish (left). A vertical bar represents the extent of the sequence generated in each species and genes are shown as bars (red = genes confirmed by cDNA, blue = predicted genes, green = pseudogene). Horizontal black lines link predicted orthologous genes. The names of orthologous genes identified in all three species are given in red.*

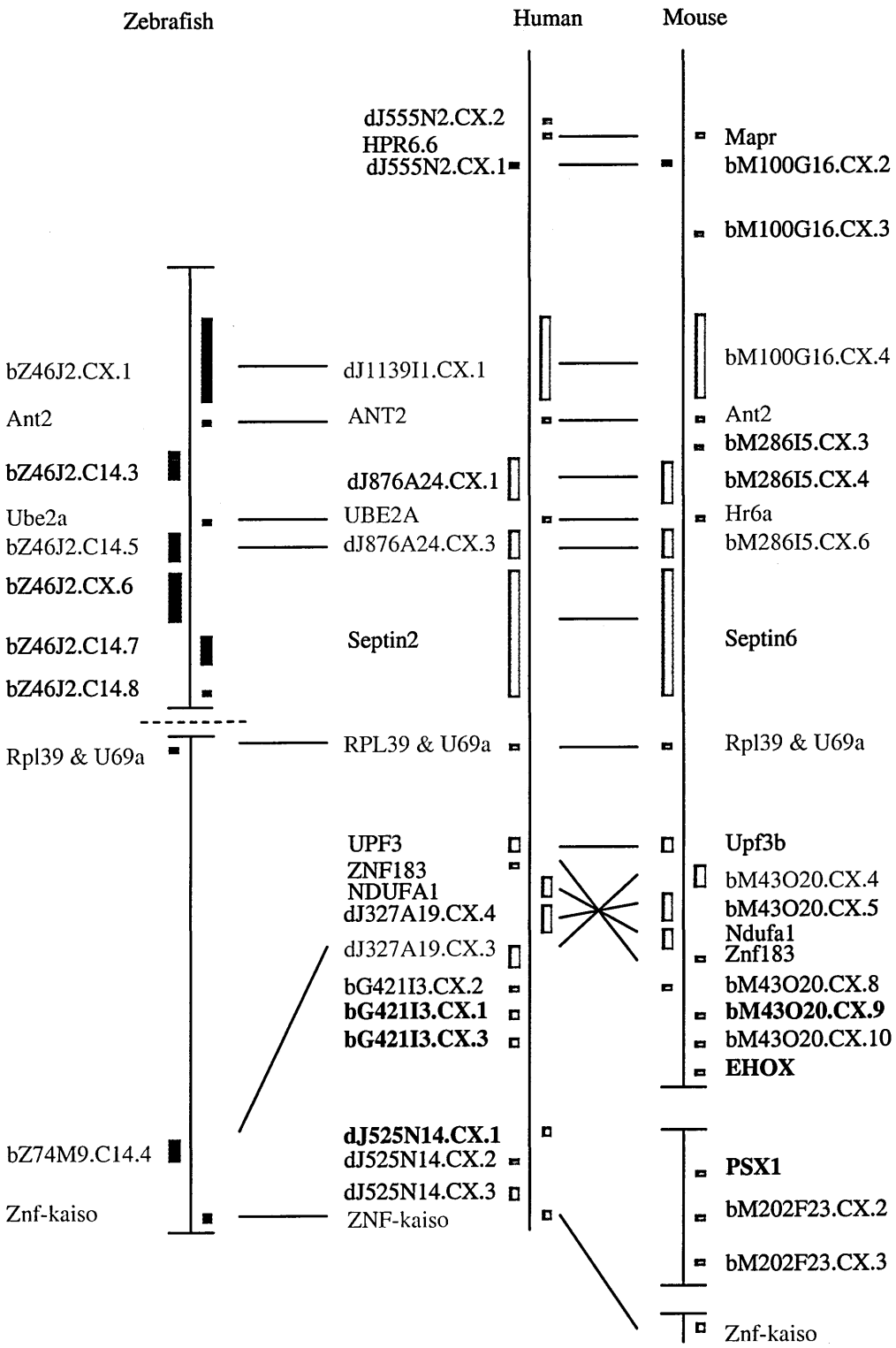

Zebrafish
Human
Mouse
dJ555N2.CX.2
HPR6.6
dJ555N2.CX.1
Mapr
bM100G16.CX.2
bM100G16.CX.3
bM100G16.CX.4
bZ46J2.CX.1
dJ1139I1.CX.1
Ant2
ANT2
Ant2
bM286I5.CX.3
bZ46J2.C14.3
dJ876A24.CX.1
bM286I5.CX.4
Ube2a
UBE2A
Hr6a
bZ46J2.C14.5
dJ876A24.CX.3
bM286I5.CX.6
bZ46J2.CX.6
bZ46J2.C14.7
Septin2
Septin6
bZ46J2.C14.8
Rpl39 & U69a
RPL39 & U69a
Rpl39 & U69a
UPF3
Upf3b
ZNF183
NDUFA1
dJ327A19.CX.4
bM43O20.CX.4
bM43O20.CX.5
Ndufa1
dJ327A19.CX.3
Znf183
bG421I3.CX.2
bM43O20.CX.8
bG421I3.CX.1
bM43O20.CX.9
bG421I3.CX.3
bM43O20.CX.10
EHOX
dJ525N14.CX.1
bZ74M9.C14.4
dJ525N14.CX.2
PSX1
dJ525N14.CX.3
Znf-kaiso
ZNF-kaiso
bM202F23.CX.2
bM202F23.CX.3
Znf-kaiso

Figure 6.12: *(see over) Identification of conserved sequences. A schematic of the region in human between HPR6.6 and ZNF-Kaiso. A scale indicates the size of the region, and genes are shown as vertical lines or boxes (exons) linked by horizontal lines (introns). Genes transcribed on the plus strand are positioned above the horizontal line, and those transcribed on the minus strand are positioned below the line. The zebrafish orthologous counterpart has been identified for the genes shown in red. No orthologue has been identified for the genes shown in green. Each red exon indicates a region conserved in human, mouse and zebrafish. The results of three methods for identifying conserved sequences between human and mouse (discussed in chapter 5) and between human and zebrafish are shown. Red vertical lines/boxes indicate the identification of a known conserved sequence. Other coloured lines/boxes indicated the position of a novel conserved sequence predicted by either PIPMAKER (black), VISTA (blue) or BLAST (green). The position of the novel conserved sequence predicted by all three methods is shown by a dotted arrow.*

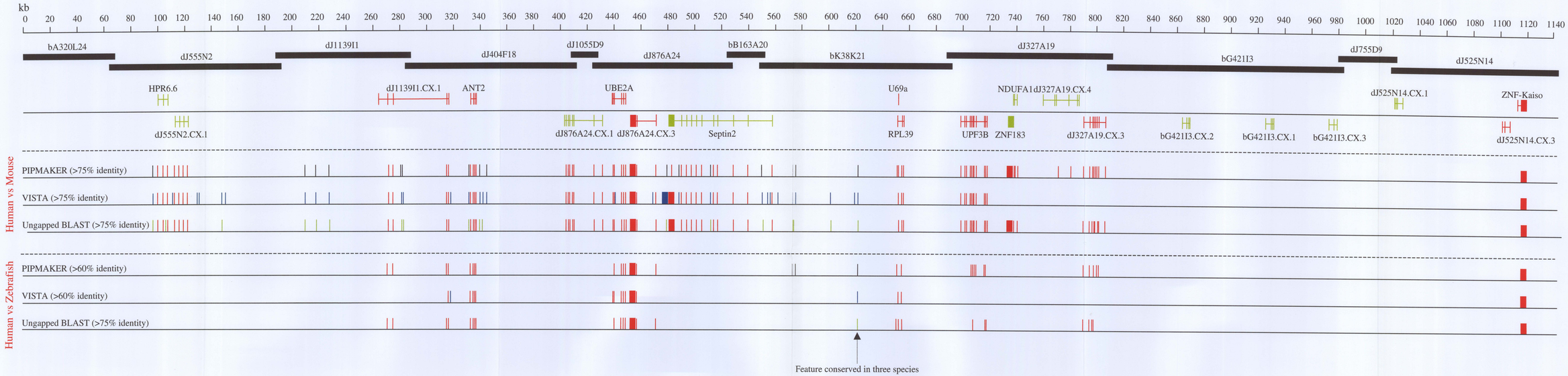

kb
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500 520 540 560 580 600 620 640 660 680 700 720 740 760 780 800 820 840 860 880 900 920 940 960 980 1000 1020 1040 1060 1080 1100 1120 1140
bA320L24
dJ555N2
dJ1139I1
dJ404F18
dJ1055D9
dJ876A24
bB163A20
bK38K21
dJ327A19
bG421I3
dJ755D9
dJ525N14
HPR6.6
dJ1139I1.CX.1
ANT2
UBE2A
U69a
NDUFA1
dJ327A19.CX.4
dJ525N14.CX.1
ZNF-Kaiso
dJ555N2.CX.1
dJ876A24.CX.1
dJ876A24.CX.3
Septin2
RPL39
UPF3B
ZNF183
dJ327A19.CX.3
bG421I3.CX.2
bG421I3.CX.1
bG421I3.CX.3
dJ525N14.CX.3
Human vs Mouse
PIPMAKER (>75% identity)
VISTA (>75% identity)
Ungapped BLAST (>75% identity)
Human vs Zebrafish
PIPMAKER (>60% identity)
VISTA (>60% identity)
Ungapped BLAST (>75% identity)
Feature conserved in three species

Table 6.4: *Summary of prediction of conserved sequences in human, mouse and zebrafish*

Method	Conserved sequence	Other sequences	Total sequences	Sensitivity	Specificity
PIPMAKER	29	3	32	0.69	0.90
VISTA	14	2	16	0.33	0.87
BLAST	26	1	27	0.62	0.96
Total	42	4	34	0.71	0.88

There are a total of 42 known sequences conserved between human, mouse and zebrafish, which are the exons of the orthologous genes. PIPMAKER identified 29 of the 42 (69%), VISTA identified 14 (33%) and BLAST identified 26 (62%). In general, fewer conserved sequences were identified in the human-zebrafish comparison than with the human-mouse comparison because the percentage identity between human and zebrafish exons was much lower. Where the percentage identity remained high, the conserved sequences were identified. This was seen for the ANT2 gene where all three methods identified all four exons and the ANT2 gene is greater than 90% identical at the nucleotide level for the coding region between all three species. In contrast, the first coding exon of the human gene dJ327A19.CX.3 is 86% identical to the mouse orthologue bM43O20.CX.4, but only 48% identical to the zebrafish orthologue bZ74M9.C14.4, and the match was not detected in zebrafish by the methods used.

The specificity of the three methods increased significantly when comparing human and zebrafish sequence as opposed to human and mouse sequence. The specificity is calculated assuming that all of the novel conserved sequences between human and mouse, predicted by the three methods in the previous chapter that lay outside the known coding sequences, were false. PIPMAKER predicted only three novel conserved sequences, VISTA predicted two and BLAST predicted only one. One

novel conserved sequence was predicted by all three methods (indicated in Figure 6.12), and further analysis of the region containing this feature in the human sequence showed that it lay within a predicted CpG island (predicted by CpGfinder). There was also a match to a human 5' EST sequence (Em:BG118506) that showed no evidence of splicing, and no apparent polyadenylation signal (see Figure 6.13). Previous analysis of predicted genes in this region (discussed in chapter 4) excluded this feature as a potential gene, as the supporting evidence was not sufficient to follow up given the guidelines outlined in Section 4.2. However, the apparent conservation of the feature in human, mouse and zebrafish, provides more confidence that the feature may be functional. In an attempt to determine whether this does represent novel coding sequence in the region, an STS, stbK38K21.3 was designed and used to screen DNA pools representing 12 different human cDNA libraries (v1-12 see Section 2.8.3). No expression was detected in any of the available cDNA libraries. However, given the amount of supporting evidence for this region of sequence, it is likely that a novel functional unit has been identified.

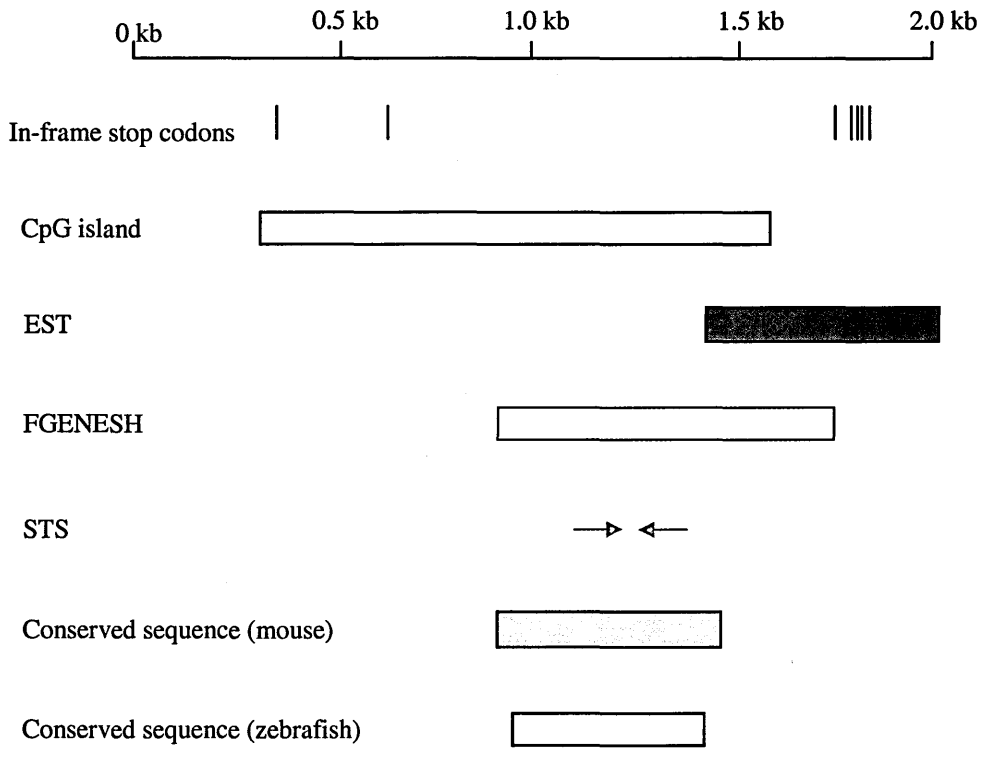


Figure 6.13: *Evidence of a novel conserved exon. Mouse DNA homology (green box) and zebrafish DNA homology (red box) positioned in same sequence as CpG island (yellow box), non-splicing EST (purple box) and FGENESH prediction (white box). The position of in-frame stop codons, based on the FGENESH prediction (shown as vertical black lines) delineates an ORF of 1.1 kb. An STS, stbK38K21.3 designed to part of the ORF (primers shown as red arrows) failed to identify any positive pools in the human cDNA libraries currently available.*

6.8 Discussion

Zebrafish bacterial clone isolation has been carried out using probes generated from STSs designed to human exons to screen a zebrafish BAC library by reduced stringency hybridisation. Analysis of the sequence of all the clones identified by reduced stringency hybridisation showed that only two of the fourteen clones contained orthologous sequences to the region of interest between HPR6.6 and ZNF-Kaiso in human. Evaluation of the method revealed two limitations. Firstly, a number of false positives were identified which did not appear to contain any sequence homologous to the human-specific sequence from which the probe was derived. Increasing the stringency of the washing after the hybridisation reduced the number of false positive clones identified, but also increased the risk of generating false negatives. Secondly, the method was not sufficiently sensitive to detect sequences less than 75% identical. Probes derived from STSs designed to four human genes did not identify clones that were later shown to contain the orthologous gene.

Recent progress has been made in both the sequencing of zebrafish ESTs, and the positioning of them within the zebrafish genome by RH mapping (Johnson, unpublished, see http://zfish.wustl.edu). Analysis of the available EST sequence data revealed two ESTs, localised to the same region on LG14 of the zebrafish genome, that are orthologous to two human genes in the region of interest between HPR6.6 and ZNF-Kaiso. This analysis reveals that the syntenic portion of part of the region of interest in human between dJ1139I1.CX.1 and ZNK-Kaiso is located on LG14 in zebrafish.

The zebrafish EST sequence data provides a more reliable method for the generation of bacterial clone contigs covering regions syntenic to human in the zebrafish and the strategy relies on the identification of orthologous gene sequences between the two organisms. An STS assay can be designed using this sequence and used to produce a probe that can then be labelled and hybridised to gridded arrays of zebrafish bacterial clones. Despite its higher accuracy, this approach is obviously limited by the availability of zebrafish specific cDNA or EST sequences. So in this case, only two human genes could be used to identify clones in zebrafish, as orthologous zebrafish EST sequences were only identified for dJ876A24.CX.1 and dJ327A19.CX.3. This method is applicable to comparative analysis between any two organisms for which orthologous sequences have been identified and is currently being applied to identify zebrafish BAC clones containing sequences orthologous to human chromosome seven (E. Green unpublished).

The two clones isolated by reduced stringency hybridisation, bZ46J2 and bZ74M9, appear to represent a region in zebrafish that is syntenic to a portion of human Xq24 between HPR6.6 and ZNF183. The evidence is based on the identification of zebrafish genes that are predicted to be the orthologues of eight of the human genes. A further four genes were identified in bZ46J2 and bZ74M9, three of which show similarity at the protein level to genes in human 11q13, suggesting a possible novel syntenic block.

A combination of *de novo* gene prediction and similarity searches predicted twelve genes in bZ46J2 and bZ74M9. Carrying out this type of analysis to identify genes in zebrafish is more difficult than in human and mouse, the other organisms studied here. The twelve genes identified in this study in zebrafish could not be confirmed by publicly availably cDNA sequence (as was the case for some of the genes described in chapter 4), and only had partial confirmation by EST sequence from zebrafish. In the cases where genes are predicted based on their similarity to sequences (both cDNA, protein and genomic) from distantly related organisms it can be very difficult to identify the exact exon boundaries as the level of similarity can be low (50-60% identity).

The region in human between HPR6.6 and ZNF-Kaiso has been compared to that in mouse and zebrafish. The availability of sequences in human, mouse and zebrafish thought to have descended from the same region in a common ancestor allows for analysis of the sequences conserved between them. Sequences that have maintained the same function in all three species are likely to be conserved. Comparisons of the conserved sequences identified in this chapter and chapter 5 shows that human and zebrafish sequences are mainly conserved in regions predicted to be coding. This compares to human and mouse sequences which show conservation outside the coding regions. The presence of a sequence conserved in all three organisms in the region between HPR6.6 and ZNF-Kaiso suggests the presence of a novel functional element. One possibility is that it is a novel exon and further screening in a wider variety of cDNA libraries (derived from human, mouse and zebrafish resources) may identify a cDNA clone to confirm this.

Multiple sequence analysis carried out in this chapter has utilised three methods, PIPMAKER, VISTA and BLAST. However, in reality the analysis that has actually been performed is a series of pairwise comparisons with the data being presented in a single view. This does not take into account conservation in the three species directly, although the results from the pairwise comparisons were considered together manually (as shown in Figure 6.12). It is not currently possible to carry out this type of analysis automatically. The sequence of the human genome is nearing completion and the sequencing of both the mouse and the zebrafish genome is underway. The sequencing of other genomes such as *S. cerevisiae* and *D. melanogaster* are available and sequencing of other vertebrate genomes is being discussed. Improvements in the tools to analyse these long sequences are required in order to extract the maximum amount of information from comparative sequence analysis.

6.9 Appendix

Table 6.5: *Comparison of orthologous genes in human (left), mouse (middle) and zebrafish (right) (a '+' indicates UTR that spans multiple exons)*

Gene Name	dJ1139I1.CX.1	bM100G16.CX.4	bZ46J2.C14.1
5'UTR	0	0	75
exon 1	152	275	272
2	242	242	242
3	173	173	173
4	135	135	135
5	201	57	212
3'UTR	351	464	0
Total Coding	903	882	1034
Total cDNA size	1254	1346	1109
intron 1	6781	8114	5105
2	3488	5393	97
3	41646	17541	984
4	721	1264	990
Total Intron size	52636	32312	7176
Genomic coverage	53890	33658	8285
Gene Name	**ANT2**	**Ant2**	**Ant2**
5'UTR	70	85	87
exon 1	111	111	111
2	487	487	487
3	141	141	141
4	158	158	158
3'UTR	258	263	305
Total Coding	897	897	897
Total cDNA size	1225	1245	984

intron 1	1034	1001	1015
2	225	406	80
3	387	510	84
Total Intron size	1646	1917	1179
Genomic coverage	2871	3162	2163
Gene Name	**UBE2A**	**Hr6a**	**bZ46J2.C14.4**
5'UTR	174	128	124
exon 1	44	44	44
2	81	81	81
3	26	26	26
4	90	90	90
5	89	89	89
6	129	129	129
3'UTR	1162	1075	14
Total Coding	459	459	459
Total cDNA size	1795	1662	597
intron 1	145	181	219
2	393	375	401
3	6106	66464	2323
4	991	842	2278
5	450	324	92
Total Intron size	8085	68186	5313
Genomic coverage	9880	69848	5910
Gene Name	**dJ876A24.CX.3**	**bM286I5.CX.6**	**bZ46J2.C14.5**
5'UTR	690+34	108+34	160+129+6
exon 1	109	97	2220
2	1958	1958	
3'UTR	1015	1015	418
Total Coding	2067	2055	2220

Total cDNA size	3082	3070	2638
intron 1	13217	10980	180
2	1061	1112	3321
Total Intron size	14278	12092	3501
Genomic coverage	17360	15162	6139
Gene Name	**RPL39**	**Rpl39**	**bZ74M9.C14.1**
5'UTR	67	275	-
exon 1	3	3	3
2	104	104	104
3	49	49	49
3'UTR	178	181	-
Total Coding	156	156	156
Total cDNA size	401	612	156
intron 1	1562	1036	201
2	3175	1235	1661
Total Intron size	4737	2271	1862
Genomic coverage	5138	2883	2018
Gene Name	**U69a**	**U69a**	**bZ74M9.C14.2**
5'UTR	-	-	-
exon 1	132	132	64
3'UTR	-	-	-
Total Coding	-	-	-
Total cDNA size	132	132	64
Genomic coverage	132	132	64
Gene Name	**dJ327A19.CX.3**	**bM43O20.CX.4**	**bZ74M9.C14.4**
5'UTR	38	184	262
exon 1	386	380	308
2	81	81	81

3	71	71	64
4	135	141	189
5	64	64	150
6	110	110	175
7	76	76	-
8	150	150	-
9	175	175	-
3'UTR	161	317	276
Total Coding	1248	1248	968
Total cDNA size	1447	1749	1506
intron 1	4409	3602	887
2	2047	5453	541
3	180	169	3319
4	1758	2784	93
5	2277	1897	625
6	74	89	
7	1770	992	
8	4621	4736	
Total Intron size	17136	19722	5465
Genomic coverage	18583	21471	6971
Gene Name	**ZNF-kaiso**	**Znf-kaiso**	**bZ74M9.C14.6**
5'UTR	134+2	185+2	0
exon 1	2019	2016	1875
3'UTR	321	453	0
Total Coding	2019	2016	1875
Total cDNA size	2340	2469	1875
intron 1	2427	2132	0
Total Intron size	2427	2132	0
Genomic coverage	4767	4148	1875

Chapter 7

Discussion

7.1 Advances in mapping technology and strategy

The strategy for mapping the human genome was developed using the methods and experience gained from mapping the genomes of model organisms. Bacterial clone maps covering the genomes of both *C. elegans* and *S. cerevisiae* were constructed by restriction digest fingerprinting whole genomic libraries of clones, which were assembled into contigs. For *C. elegans*, once the bacterial clone resources had been exhausted, the remaining gaps were bridged using YACs. The increased complexity of the human genome (30 times larger than the *C. elegans* genome) meant that the strategy implemented for mapping the *C. elegans* genome could not be directly applied to the human genome. For instance, whole genome fingerprinting of a human cosmid library was not feasible given the increased size of the human genome. This increase in genome complexity would have meant a large increase in the number of cosmids required (approaching half a million cosmids for a six fold coverage of the human genome), which in turn would have made the interactive contig assembly in FPC too laborious.

By contrast, strategies for generating clone maps covering small regions of the human genome relied on the use of YACs from the outset. Mapping regions of the human genome involved the use of landmarks taken from previously published genetic maps or RH maps which were used to identify and order YAC clones for contig construction. This approach was scaled up to construct maps on some chromosomes but attempts to adapt the whole genome fingerprinting strategy employed for the model genomes failed because of the instability and chimaerisms that was present in the YACs. YACs were also considered to be inappropriate substrates for sequencing,

compared to bacterial clones. As a result, work on the human genome was concentrated on finding ways to construct bacterial clone maps. Early on, cosmid contigs were constructed using data from the available YAC contigs and associated landmarks. For example, whole YACs were hybridised to gridded cosmid libraries to identify underlying cosmids. The cosmids were then fingerprinted to assemble contigs using the same method of restriction digest fingerprinting that was developed for the *C. elegans* mapping project. In one case, this approach resulted in the assembly of contigs covering 80% of the 450 kb region, and the bacterial clone contigs were bridged by the starting YAC (Holland, J., *et al.*, 1993). This integrated YAC-cosmid map therefore closely resembled the product of the *C. elegans* mapping effort, although it was generated in a different way. This approach was initially employed as the basis for constructing the Xq22 contig described in chapter 3, in which cosmids were identified using probes derived both from YACs and from the available landmarks ordered in the YAC contig. Cosmids were then assembled into contigs using restriction digest fingerprinting. The progress was slow given the small size of cosmids (40 kb) and resulted in only 50% coverage of the 6.5 Mb region in cosmid contigs.

A major advance in human genome mapping was the development of two larger insert bacterial-based cloning systems (PACs and BACs) with the combined advantages of a much larger insert size (up to 300 kb), and little or no chimaerism or instability in contrast to that seen in YACs. This development was reflected in phase 2 of the Xq22 project, in which the focus of the mapping switched to PAC identification. Cosmids at the ends of contigs were labelled and used as hybridisation probes to isolate PACs. The PACs were fingerprinted and incorporated into the existing contigs. Difficulties

arose in using a combination of small and large clones together in terms of both fingerprint comparisons and in the selection of a minimum set of clones for sequencing. For fingerprinting, false overlaps between PACs were observed and true overlaps between PACs and cosmids were missed; for minimum set selection, the large insert clones made the cosmids largely redundant. Increasingly, sequence-ready contigs were constructed in new regions of the genome using only the larger insert PACs and BACs, including the region discussed in chapter 4, where contigs were constructed as part of the whole X chromosome mapping project.

At this stage in the human genome mapping project, bacterial clone contig construction was still reliant on landmarks ordered on available YAC contigs. On some chromosomes, including chromosome 22 and the X chromosome, the continued development of YAC contigs resulted in the provision of a sufficiently high density of marker (greater than 1 per 70 kb for chromosome 22 – Collins, J. E. *et al*, 1995) to enable construction of PAC and BAC contigs directly. This was illustrated in phase 3 of the Xq22 project, where new landmarks which became available from detailed YAC maps were used to identify new PACs. Sufficient coverage of the region was obtained to allow closure by walking in the final phase (phase 4) of the Xq22 map.

YAC maps, or later RH maps, of similar quality provided a high density of ordered landmarks along each chromosome and provided the means to apply the same strategy to map the rest of the human genome. However, the increased insert size of BACs and PACs compared to cosmids led to a further development late in the project. Whole genome fingerprinting, which had been used successfully to map the *C. elegans* genome in cosmids, was introduced using BACs to obtain very rapid coverage of the

rest of the human genome independently of landmarks, thus reducing the number of landmarks required to anchor and orient the bacterial clone map.

Walking to close gaps in the human genome was aided by the generation of sequence at the ends of BAC clones. These sequences provided a large resource of new landmarks both within contigs to confirm fingerprint assemblies but also at the ends of contigs to screen for additional BACs for gap closure. Bacterial clone maps now cover most of the euchromatic portion of the human genome, and the remaining gaps are being closed. An interesting development is the return to YACs to bridge gaps for which there is no bacterial clone coverage, mimicking the final gap closure carried out in the *C. elegans* mapping project.

The advances in bacterial clone mapping and sequencing that evolved for the construction of the sequence-ready maps for the human genome are now being applied to other organisms, particularly mouse and zebrafish. For mouse, whole genome fingerprinting and assembly of mouse-derived bacterial clones generated 7,500 contigs which were extended and joined to form less than 400 contigs covering approximately 90% of the mouse genome. Mouse BAC end sequences were used to align the mouse contigs to the human genome sequence and accelerated the manual joining process (Gregory, S., unpublished, http://mouse.ensembl.org/). A similar project is now well underway to generate bacterial clone contig maps of the zebrafish genome (see http://www.sanger.ac.uk/Projects/D_rerio). The speed and accuracy with which these large genomes are being mapped is a reflection of the technologies applied to map the human genome sequence. The ability to use the human genome sequence as a framework to anchor the mouse bacterial clone maps takes advantage of

the sequences conserved between the two genomes. The mapping of both the mouse and zebrafish genomes does not rely solely on using the human genome as a framework. Independent analysis of data, using whole genome fingerprinting is ensuring that the species-specific organisation of the mouse and zebrafish genomes will be maintained. However, the ability to compare the maps and sequences with the human map and sequence allows the syntenic relationships to be established and further characterised.

The developments in mapping, including large insert bacterial cloning systems and high density landmark generation/ordering, have made large-scale map construction a very efficient process and may lead to the construction of maps covering other genomes. These maps could be used as reagents to generate high quality sequence as was seen for the human genome. Alternatively the maps could be used as frameworks to carry out either clone-based or whole genome shotgun sequencing without finishing, using the clone maps to anchor and orient the shotgun sequence. The quality of this draft product produced would be much greater than from whole genome shotgun alone. The amount of large-scale sequencing of other genomes will depend in part on the contribution the maps and sequence currently being generated in mouse and zebrafish play in interpreting the human genome, the costs involved and the value to the research of the other organisms.

7.2 Mining the human genome sequence

The complexity of an organism was always thought to be a reflection of the complexity of the gene content. Analysis of the sequence of *S. cerevisiae*, a single-celled organism predicted 6,000 genes. The sequence of the genome of *C. elegans*, a multi-cellular organism, was predicted to contain19,000 genes. However, when the sequence of *D. melanogaster* was produced, less than 14,000 genes were predicted. It became clear that gene number alone was not a direct indication of complexity. Complexity may arise from not only the number of genes, but also from the use of promoters and regulatory elements, post-transcriptional modification where alternative splicing produces multiple transcripts per gene, and protein diversity such as differences in structure and interaction. For example, early analysis of the sequence of *D. melanogaster* predicted 13,601 genes would produce at least 14,113 transcripts but that this was likely to be an underestimate (Adams, M.D., *et al*, 2000). Analysis of part of the draft sequence of the human genome, using chromosome 22 and chromosome 19 sequence and annotation, predicted an average of 2.6 (for chr22) and 3.2 (for chr19) transcripts per gene (Lander, E.S., *et al*, 2001). Mining the human genome, arguably the most complex organism of all is significantly more difficult. At the same time, virtually all the information which gives rise to the greater complexity of protein function by the mechanisms described above is nevertheless encoded within the DNA sequence of the human genome. The challenge is to find those features and understand their role.

Prior to the availability of genomic sequence, gene identification relied upon the isolation of cDNA clones using YACs or bacterial clones from the physical maps.

Physical maps were generated across regions thought to contain genes involved in particular diseases. Candidate genes in the region were identified using techniques such as cDNA direct selection and exon trapping. However, these methods were laborious and subject to artefacts. For example, cDNA direct selection suffered from false positives due to repeats, pseudogenes and gene families. The techniques also yielded significant levels of false positives, often only providing limited levels of enrichment for the cDNAs of interest, resulting in the analysis of a larger number of cDNA clones than was desirable. A significant advantage of exon trapping was that it did not rely on expressed sequences to identify genes, but used the signals encoded in the genomic sequence to identify regions that splice. Exon trapping is also prone to false positives and because it relies on the presence of flanking intron sequence, was not able to clone first or last exons, or single exon genes. Some development of the initial procedure does allow for the cloning of 5' and 3' exons but this was also prone to false positives. However, exon trapping has been applied on a large scale to identify 6,400 potential exons on chromosome 22, and estimates suggested over half represented true exons (Trofatter, J. A., *et al.*, 1995).

The availability of the genomic sequence provides the foundation for a full investigation into the features contained within it, including the genes and regulatory elements. Moreover, this analysis can be carried out in part computationally, making the process less labour-intensive. Genes can be predicted by two complementary methods, similarity searches and *de novo* gene prediction, and confirmed where necessary by generating novel cDNA sequence. The ability to predict the genes in a given region which can then be investigated by experimental methods is a much more efficient system for gene identification than was previously seen prior to the

availability of genomic sequence. This strategy for gene identification was described in chapter 4, but revealed several limitations to identifying genes in this manner.

The first limitation is the incompleteness of the genomic sequence. A complete transcript map cannot be constructed if gaps remain in both the map and sequence. This was the case in the Xq23-q24 region studied where eleven gaps were present between twelve sequence contigs, and for a small number of clones only draft sequence was available. In order to carry out a detailed analysis of the genes contained within a given region, it is important to generate high quality finished sequence covering as much of the region of interest as possible. This was highlighted in the analysis of the critical region for MRX23 for which the bacterial clone map contained a gap of approximately 500 kb. A complete analysis of the genes in the region was not possible until this sequence becomes available.

A second limitation is the lack of available supporting evidence for predicted gene structures in the DNA and protein sequence database (e.g EMBL and Swissprot). In order to confirm a predicted gene, human cDNA sequence covering at least the predicted ORF is required. Also, supporting evidence may be present for part of the gene structure and may be incomplete. For instance, a gene having a large mRNA may be incompletely reverse transcribed to form a partial cDNA clone or the cDNA clone is incompletely sequenced as is the case in EST sequencing. These problems are being addressed in part by the full length cDNA sequencing programs such as the Mammalian Gene Collection (MGC - http://mgc.nci.nih.gov/). However, in many cases it is still necessary to carry out targeted cDNA sequencing to confirm predicted gene structures. Recent estimates from the genes identified on chromosome 20

showed that of the 727 genes identified, 350 need additional confirmatory human cDNA sequence (Graeme Bethel, personal communication). A third limitation in gene identification is the inability to confirm predicted genes. Some genes remain unconfirmed as they are not represented in the cDNA libraries available. They may be expressed in tissues not represented in the library collection, or temporally or transiently expressed. Testing a wider variety of cDNA libraries will increase the likelihood of confirming the gene. Predicting the mouse orthologue and testing for the presence of the cDNA in mouse cDNA libraries provides more flexibility in terms of the range of tissues that can be utilised and this approach is already providing valuable information for human sequence annotation. For example, the RIKEN Institute are sequencing cDNA clones, derived from over 200 different tissues and cell types, with the aim of collecting data on as many full-length cDNAs as possible (http://genome.gsc.riken.go.jp). In conjunction with the human cDNA sequencing effort, the Mammalian Gene Collection (MGC) is also sequencing cDNA clones derived from mouse tissues (http://mgc.nci.nih.gov/). As an alternative to cDNA analysis, the screening of RNA by RT-PCR may identify transcripts not previously detected as the RNA has undergone fewer manipulations. Very low level transcripts have also been detected by nested RT-PCR in tissues not expected to express the gene and may represent illegitimate transcription which nevertheless confirms the biological activity of the predicted gene (Roberts, R. G., *et al.*, 1991; Kaplan, J. C., *et al.*, 1992).

It is unlikely that all genes will be confirmed by identification and sequencing of cDNA. Therefore, the predicted genes for which no cDNA sequence is available require the level of confidence associated with each prediction to be assessed. In

chapter 4, predicted genes were only followed up if there was a certain level of supporting evidence, such as regions predicted by two separate gene prediction programmes, or an EST or cDNA sequence that spliced exactly on to the genomic DNA. Given the need to set a series of criteria by which predicted genes were analysed in order to reduce over prediction of genes to a minimum, it is possible that some real genes are missed. Therefore it is important to generate as much information as possible to be confident that a high percentage of the real genes have been identified, whilst limiting the number of false predictions. In the absence of supporting cDNA or protein sequence, the best indication that predicted genes or exons from partially confirmed genes are real is the observation that the predicted gene is conserved in other species as was carried out in chapters 5 and 6 (see Section 7.4).

As discussed in chapter 4, the first exon of a gene, containing the 5' untranslated region is often the most difficult to identify. The lack of coding sequence limits the use of sequence similarity searches. In addition, the gene prediction programs are not designed to predict UTR, because the prediction process involves codon usage analysis. Given the ever- increasing amount of gene structure information being generated, it may be possible to design computational tools to specifically predict first exons. In a similar fashion to other software development, a high quality training set of known first exons could be scanned and compared to identify signals specific to first exons. The prediction program could be tested and improved using a second calibration set.

The first exon of a gene is often adjacent to the promoter elements and transcription start site and analysis of these features may aid in the identification of the first exon. Three methods for the analysis of 5' sequence elements are described in chapter 4: CpG island detection, PromotorInspector and Eponine, which begin to address this. However, the analysis is currently limited to genes which are associated with CpG rich sequences at their 5' end. These tools are continuing to improve and the development could be widened to include genes which are not associated with a CpG island.

The manually curated annotation of the genomic sequence described in chapter 4 neither confirms all of the predictions, or identifies all of the functional elements contained within the sequence but still goes further than most of the automatic human sequence analysis currently being carried out (e.g genome browsers such as ENSEMBL and UCSC). The majority of the genes identified by the methods described in chapters 4, 5, and 6 are protein-coding genes, because these are the ones that can be more readily identified by primary sequence analysis. However, other genes produce non-coding RNAs (ncRNA) that function directly as structural, regulatory or even catalytic RNAs (Eddy '99). Unlike protein-coding genes these ncRNA genes have no obvious primary sequence patterns that can be used to identify them. Separate tools need to be developed to identify these sequences. One strategy that may be applicable to human sequence has been used successfully to identify potential novel ncRNAs in *E. coli*. The method took regions conserved in multiple species of bacteria and identified candidate ncRNA based on secondary structure analysis, such as identification of hairpin loops, rather than primary sequence analysis.

7.3 Comparing different genomes to aid human genome sequence analysis

Comparing genome sequences from different species is a powerful method for increasing the confidence in predicted genes, or identifying novel functional units. When two species diverge from a common ancestor those sequences that maintain their original function are likely to remain conserved in both species throughout their subsequent divergent evolution. In order to identify the functionally important units in the human genome it may be necessary to compare genome sequences from a variety of organisms, although any human-specific features will not be detected by this strategy. The more distantly related organisms such as yeast and worm are likely to show sequence conservation in coding regions alone. This may also be the case for distantly related vertebrates such as fish. The more closely related organisms, such as mouse, are likely to be conserved in coding regions, but also in other functional elements such as regulatory sequences. However, the closer the evolutionary relationship with human, the more 'sequence noise' is likely to arise where non-functional sequence appears similar because insufficient time has elapsed for the two sequences to diverge. For instance, comparing sequence between human and chimpanzee is only really useful to determine the differences between the two species rather than the similarities. The extreme case of this is to compare genomes within the same species to identify variation within a population and to determine the functional significance of the variation.

Identification of conserved sequences outside regions predicted to be coding may indicate the presence of a novel gene or a region involved in gene regulation. The challenge when comparing genome sequences is to identify the sequences conserved

between different organisms given that many are short and interspersed with large regions of non-conserved, non-functional DNA, and to distinguish between the conserved sequences that are functional and those that have no function. The coding regions of many vertebrate genes, unlike genes of other lower organisms such as bacteria and most of the genes in yeast, are disrupted by introns and exons can be small. Other functionally important segments of DNA such as regulatory sequences can be very small and placed far away from the gene they influence. In chapters 5 and 6, a genomic region in Xq24 was analysed in mouse and zebrafish, to identify potential novel functional elements using comparative sequence analysis.

There is a range of comparative sequence analysis tools available for identifying sequences conserved between species, some of which are described and used in chapters 5 and 6. Generating either local alignments or global alignments between pairs of sequences identifies conserved sequences between different species. Three tools, PIPMAKER, VISTA and BLAST, were used in chapters 5 and 6 to identify conserved sequences between human and mouse, human and zebrafish, and sequences conserved in all three species, but further work needs to be carried out to determine whether or not these novel conserved sequences represent functional units, for example involved in gene regulation.

Experiments have been carried out to identify those conserved elements that have regulatory function but these have been carried out on individual regions of interest and not on a large scale. In the case of the SCL locus, a region conserved in human and mouse was shown using a transgenic xenopus assay to be a novel neural enhancer (Gottgens, B., *et al.*, 2000). Touchman, J. W., *et al.* (2001) identified a novel

regulatory element of the human and mouse α-Synuclein genes using comparative genome sequence analysis and subsequently confirmed its regulatory function using a reporter gene assay (Touchman, J. W., *et al.*, 2001). Other methods of functional testing include mobility shift assays and DNA footprinting, two methods that examine the binding of proteins to the DNA fragments of interest, and site-directed mutagenesis, followed by functional analysis can identify individual bases that affect regulation (Sambrook, J., *et al.*, 1989).

7.4 Functional analysis of gene products

The identification of the genes encoded within the human genome is only the first step to a complete understanding of the genes encoded in the human genome. An important second step is to understand the role these genes play in the correct functioning of a cell. One method for determining function is to see the phenotypic effect if the function of a gene is altered. In humans, this analysis has been provided by naturally occuring mutations that cause diseases. Cross-species comparison is also a powerful tool for gaining an insight into the function of a gene where the function of one gene can be inferred based on the function of the homologue. The advantage of functional analysis in other organisms is that these mutations can be engineered. This was first pioneered with mutation analysis in bacteria and drosophila as large numbers of mutant phenotypes can be analysed.

More recently chemical mutagenesis programmes have been undertaken in yeast, worm, zebrafish and mouse (reviewed in Justice, M. J., 2000) where alterations in

phenotype are being used to categorise the function of genes that have been mutated. Genetic screens in the zebrafish have proven particularly useful in identifying genes involved in development. Zebrafish eggs develop externally and they can be easily visualised. The embryos are relatively transparent which aids the detection of phenotypic abnormalities. Two studies which screened for both defects in embryogenesis and essential functions and identified more than 2000 mutant phenotypes (Driever, W., *et al.*; 1996, Haffter, P., *et al.*, 1996). The fact that embryo development in the mouse takes place in the uterus means that in mouse mutagenesis programmes these types of mutations are difficult to analyse. Two groups (UK ENU Mutagenesis programme and German ENU-mouse mutagenesis screening project) are using a breeding strategy to identify dominant and haplo-insufficient mutations. The UK programme is screening for visible phenotypes including sensory, neurological, neuromuscular alterations and behavioural assays, and the German project is screening for haematological, clinical chemistry, immunological and allergy defects. An alternative to the *in vivo* mutagenesis programs is *in vitro* mutagenesis by gene trapping, or gene targeting in mouse ES cells. These methods allow mutations to be characterised in cell culture before translation to mouse (Evans, M. J., *et al.*, 1997). Gene trapping involves random integration in to the genome of a promoter-less reporter gene construct. Incorporation of the reporter gene within a gene may abolish the function of that gene as well as enable the selection of the mutant on the basis of the transcribed reporter gene. As an alternative to random mutagenesis, it is possible to target specific genes for disruption. One of the best methods for incorporating mutations into the mouse genome has been gene targeting by homologous recombination in ES cells and modifications to this procedure enable virtually any designer modification to be introduced to known cloned mouse genes (Koller, B. H.,

et al., 1992). However, gene targeting requires some knowledge of the sequence of the gene of interest, although recent developments require only a small amount of homologous sequence (approximately 80 bp) (Zhang, P., *et al.*, 2002).

Obtaining functional information for genes in other organisms allows inferences to be drawn about the function of the orthologous genes in humans. However, inferring the function of a gene in one organism based on evidence from another organism does have limitations. As soon as speciation from a common ancestor occurs, each new species is able to evolve independently. Even though initially orthologous genes may have the same function, over time novel function for one or both genes may evolve (see Figure 7.1). For instance, duplication of DNA sequence that includes a single gene will create two copies of the orthologue in one species. The duplicated copies are now termed paralogues. Although the one paralogue may maintain its original function, the other paralogue may evolve a novel function, as there may be little or no selective pressure for it to maintain its original function. Inferring the function of the orthologous gene, based on the function of the second paralogue would be incorrect. An example of this is the mouse p53 tumour suppressor, a transcription factor that regulates the progression of the cell through its cycle and cell death in response to environmental stimuli. In contrast, p63, presumed to be a paralogue based on its sequence similarity to p53, has a completely separate function and is essential for several aspects of ectodermal differentiation during embryogenesis (Mills, A. A., *et al.*, 1999).

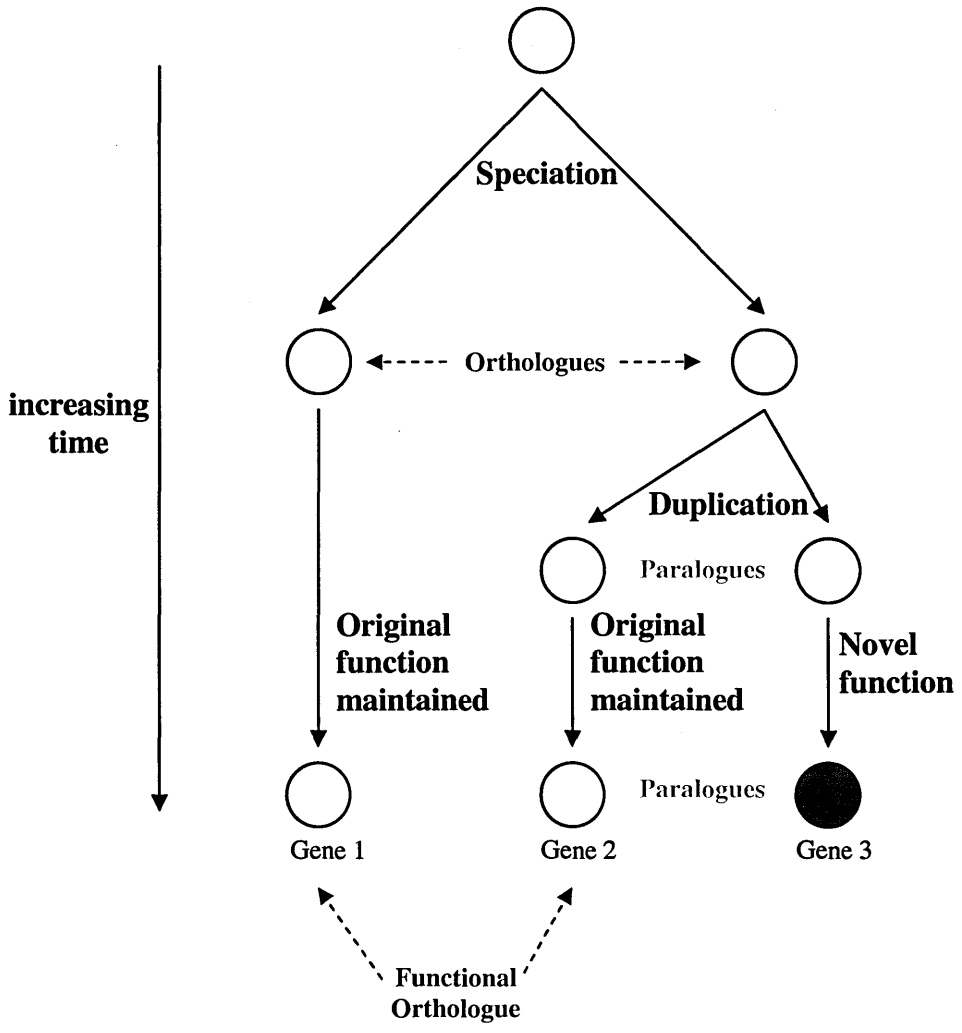


Figure 7.1: *A representation of how speciation and gene duplication can influence comparative genome analysis. The example shows a gene (circle) with a specific function (colour of circle). Two species diverge from a common ancestor (through speciation) which creates two orthologoues. Gene duplication in one species, creates two paralogues, one of which is free to evolve a novel function (represented by a brown circle). Although originally descended from the same gene, inferring function of gene 1 from the function of gene 3 would be incorrect. The true functional orthologue of Gene 1 is Gene 2, whose presence may not be detected by localised genome comparison.*

Analysis of a small region in each organism may not detect other copies of the same gene. Establishing the extent of paralogy within each organism is essential before the function of a particular gene is inferred by cross-species analysis. Detailed synteny maps such as those generated in chapters 5 and 6, and complete analysis of the potential orthologous genes are required to increase the confidence that true orthologues have been identified.

Further insight into the function of a protein can also be obtained by analysing the protein sequence for functional domains, using sequence-motif searches such as those available at INTERPRO (see chapter 4). Although these methods may be used to gain information regarding the potential function of a protein, experimental verification is always required to determine the exact function. The improvements in the speed and accuracy of gene identification have meant that an increasing number of genes have been and are being identified in a semi-automated fashion. However, little experimental data is being generated to confirm the predicted protein sequences.

Ultimately, confirmation of the structure and function of proteins is required by species-specific investigation where possible. For instance, the structure of a protein can be determined using X-ray crystallography, nuclear magnetic resonance and mass spectrometry. Determining the cellular localisation of a protein by generating and expressing a fusion construct containing the gene of interest linked to a reporter molecule such as green fluorescent protein (GFP) will provide insight into where in the cell the protein is functional. The hybrid protein is very different to the original protein of interest, therefore this type of analyses is often carried out using different reporter molecules, or antibodies to universal tags, and attaching the reporter molecule

to both the 5' and the 3' end of the protein of interest. Alternatively, an antibody specific to the protein of interest can be generated but this is laborious and costly so would be difficult to do on all human proteins.

Protein binding assays can be carried out using the yeast two hybrid and mammalian two hybrid system. In the yeast two-hybrid (Y2H) system, the gene of interest is commonly linked to the Gal4 DNA binding domain and is co-transfected into yeast cells with a library of genes linked to the Gal4 activation domain. If the protein of interest binds to a target protein, the two Gal4 domains will be brought together, and the expression of the downstream *LacZ* reporter gene can be measured using the β-galactosidase assay. One advantage of the Y2H system is that when a positive match is detected, the ORF is identified by simply sequencing the relevant clone(s). Therefore Y2H system is amenable to high-throughput screening of protein-protein interactions, such as has been reported for *S. cerevisiae* and *C. elegans* (reviewed in Walhout, A. J., *et al.*, 2000).

7.5 Conclusion

Prior to the generation of whole genome sequences, individual research focussed on understanding individual genes or gene families, their protein products and their function. The availability of large amounts of genome sequence from human and other organisms is enabling large-scale interpretation of the sequences, including all the functional elements encoded within. This has seen the growth of large-scale mapping and sequence production, along with the development of computational hardware and software to both store and analyse this data. However, the complete interpretation of the genome sequence of these organisms includes the complete characterisation the genes encoded within and the role each gene plays within a cell and the contribution to the whole organism. This characterisation will need to be carried out for each gene and each functional element in each genome. Individual research will focus on understanding individual genes, gene families, their protein products and their function.

Chapter 8

References

Aaronson, J. S., Eckman, B., Blevins, R. A., Borkowski, J. A., Myerson, J., Imran, S., *et al.* (1996). Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res* **6:** 829-45.

Abbs, S., Roberts, R. G., Mathew, C. G., Bentley, D. R. and Bobrow, M. (1990). Accurate assessment of intragenic recombination frequency within the Duchenne muscular dystrophy gene. *Genomics* **7:** 602-6.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., *et al.* (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252:** 1651-6.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3-174.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., *et al.* (2000). The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-95.

Ahmad, K. F., Engel, C. K. and Prive, G. G. (1998). Crystal structure of the BTB domain from PLZF. *Proc Natl Acad Sci U S A* **95:** 12123-8.

Albertsen, H. M., Abderrahim, H., Cann, H. M., Dausset, J., Le Paslier, D. and Cohen, D. (1990). Construction and characterization of a yeast artificial chromosome library containing seven haploid human genome equivalents. *Proc Natl Acad Sci U S A* **87:** 4256-60.

Aldridge, J., Kunkel, L., Bruns, G., Tantravahi, U., Lalande, M., Brewster, T., *et al.* (1984). A strategy to reveal high-frequency RFLPs along the human X chromosome. *Am J Hum Genet* **36:** 546-64.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-402.

Aman, M. J., Tayebi, N., Obiri, N. I., Puri, R. K., Modi, W. S. and Leonard, W. J. (1996). cDNA cloning and characterization of the human interleukin 13 receptor alpha chain. *J Biol Chem* **271:** 29265-70.

Anand, R., Riley, J. H., Butler, R., Smith, J. C. and Markham, A. F. (1990). A 3.5 genome equivalent multi access YAC library: construction, characterisation, screening and storage. *Nucleic Acids Res* **18:** 1951-6.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res* **9:** 3015-27.

Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., *et al.* (1998). Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* **8:** 29-40.

Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90:** 11995-9.

Arabidopsis Genome Initiative, The. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Arpin, M., Friederich, E., Algrain, M., Vernel, F. and Louvard, D. (1994). Functional differences between L- and T-plastin isoforms. *J Cell Biol* **127:** 1995-2008.

Au, H. C., Seo, B. B., Matsuno-Yagi, A., Yagi, T. and Scheffler, I. E. (1999). The NDUFA1 gene product (MWFE protein) is essential for activity of complex I in mammalian mitochondria. *Proc Natl Acad Sci U S A* **96:** 4354-9.

Bailey, J. A., Carrel, L., Chakravarti, A. and Eichler, E. E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A* **97:** 6634-9.

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., *et al.* (2000). The syntenic relationship of the zebrafish and human genomes. *Genome Res* **10:** 1351-8.

Bardwell, V. J. and Treisman, R. (1994). The POZ domain: a conserved protein-protein interaction motif. *Genes Dev* **8:** 1664-77.

Barr, M. L. and Bertram, E. G. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature* **163:** 676-677.

Barr, M. L. and Carr, D. H. (1961). Correlations between sex chromatin and sex chromosomes. *Acta Cytol.* **6:** 34-45.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., *et al.* (2002). The Pfam protein families database. *Nucleic Acids Res* **30:** 276-80.

Battini, R., Ferrari, S., Kaczmarek, L., Calabretta, B., Chen, S. T. and Baserga, R. (1987). Molecular cloning of a cDNA for a human ADP/ATP carrier which is growth-regulated. *J Biol Chem* **262:** 4355-9.

Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. and Lander, E. S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* **10:** 950-8.

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10:** 1001-10.

Bell, C. J., Budarf, M. L., Nieuwenhuijsen, B. W., Barnoski, B. L., Buetow, K. H., Campbell, K., *et al.* (1995). Integration of physical, breakpoint and genetic maps of chromosome 22. Localization of 587 yeast artificial chromosomes with 238 mapped markers. *Hum Mol Genet* **4:** 59-69.

Bellanne-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., Bertrand, S., *et al.* (1992). Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* **70:** 1059-68.

Belyaev, N., Keohane, A. M. and Turner, B. M. (1996). Differential underacetylation of histones H2A, H3 and H4 on the inactive X chromosome in human female cells. *Hum Genet* **97:** 573-8.

Bentley, D. R., Deloukas, P., Dunham, A., French, L., Gregory, S. G., Humphray, S. J., *et al.* (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409:** 942-3.

Bernstein, H. S., Bishop, D. F., Astrin, K. H., Kornreich, R., Eng, C. M., Sakuraba, H., *et al.* (1989). Fabry disease: six gene rearrangements and an exonic point mutation in the alpha-galactosidase gene. *J Clin Invest* **83:** 1390-9.

Bickmore, W. A. and Sumner, A. T. (1989). Mammalian chromosome banding--an expression of genome organization. *Trends Genet* **5:** 144-8.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., *et al.* (1997). The complete genome sequence of Escherichia coli K-12. *Science* **277:** 1453-74.

Boguski, M. S. (1995). Hunting for genes in computer data bases. *N Engl J Med* **333:** 645-7.

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., *et al.* (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351:** 325-9.

Bouffard, G. G., Idol, J. R., Braden, V. V., Iyer, L. M., Cunningham, A. F., Weintraub, L. A., *et al.* (1997). A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res* **7:** 673-92.

Boyd, Y., Blair, H. J., Cunliffe, P., Denny, P., Gormally, E. and Herman, G. E. (1998). Encyclopedia of the mouse genome VII. Mouse chromosome X. *Mamm Genome* **8:** S361-77.

Brenner, S. (1990). The human genome: the nature of the enterprise. *Ciba Found Symp* **149:** 6-12; discussion 12-7.

Brickner, A. G., Koop, B. F., Aronow, B. J. and Wiginton, D. A. (1999). Genomic sequence comparison of the human and mouse adenosine deaminase gene regions. *Mamm Genome* **10:** 95-101.

Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S., McCabe, V. M., *et al.* (1991). Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351:** 329-31.

Brockdorff, N. (1998). The role of Xist in X-inactivation. *Curr Opin Genet Dev* **8:** 328-33.

Bruls, T., Gyapay, G., Petit, J. L., Artiguenave, F., Vico, V., Qin, S., *et al.* (2001). A physical map of human chromosome 14. *Nature* **409:** 947-8.

Buetow, K. H., Ludwigsen, S., Scherpbier-Heddema, T., Quillen, J., Murray, J. C., Sheffield, V. C., *et al.* (1994). Human genetic map. Genome maps V. Wall chart. *Science* **265:** 2055-70.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Burglin, T. R. and Barnes, T. M. (1992). Introns in sequence tags. *Nature* **357:** 367-8.

Burke, D. T., Carle, G. F. and Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806-12.

Cabezas, D. A., Arena, J. F., Stevenson, R. E., Schwartz, C., Goldberg, S., Morales, A., *et al.* (1999). XLMR database. *Am J Med Genet* **85:** 202-5.

Caput, D., Laurent, P., Kaghad, M., Lelias, J. M., Lefort, S., Vita, N., *et al.* (1996). Cloning and characterization of a specific interleukin (IL)-13 binding protein structurally related to the IL-5 receptor alpha chain. *J Biol Chem* **271:** 16921-6.

Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., *et al.* (2001). Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* **77:** 79-90.

C. elegans Sequencing Consortium, The. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. *Science* **282:** 2012-8.

Chapman, V. M. (1986). X chromosome regulation in oogenesis and early mammalian development. Rossant, J. and Pedersen, E. D. Cambridge University Press.

Chen, J. D., Mackey, D., Fuller, H., Serravalle, S., Olsson, J. and Denton, M. J. (1989). X-linked megalocornea: close linkage to DXS87 and DXS94. *Hum Genet* **83:** 292-4.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G., *et al.* (1992a). Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359:** 380-7.

Chumakov, I. M., Le Gall, I., Billault, A., Ougen, P., Soularue, P., Guillou, S., *et al.* (1992b). Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genome library. *Nat Genet* **1:** 222-5.

Chumakov, I. M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., Guillou, S., *et al.* (1995). A YAC contig map of the human genome. *Nature* **377:** 175-297.

Clerc, P. and Avner, P. (1998). Role of the region 3' to Xist exon 6 in the counting process of X-chromosome inactivation. *Nat Genet* **19:** 249-53.

Coffey, A. J., Roberts, R. G., Green, E. D., Cole, C. G., Butler, R., Anand, R., *et al.* (1992). Construction of a 2.6-Mb contig in yeast artificial chromosomes spanning the human dystrophin gene using an STS-based approach. *Genomics* **12:** 474-84.

Coffey, A. J., Brooksbank, R. A., Brandau, O., Oohashi, T., Howell, G. R., Bye, J. M., *et al.* (1998). Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nat Genet* **20:** 129-35.

Cohen, D., Chumakov, I. and Weissenbach, J. (1993). A first-generation physical map of the human genome. *Nature* **366:** 698-701.

Collins, J. and Bruning, H. J. (1978). Plasmids useable as gene-cloning vectors in an in vitro packaging by coliphage lambda: "cosmids". *Gene* **4:** 85-107.

Collins, J. E., Cole, C. G., Smink, L. J., Garrett, C. L., Leversha, M. A., Soderlund, C. A., *et al.* (1995). A high-density YAC contig map of human chromosome 22. *Nature* **377:** 367-79.

Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986). Towards a physical map of the genome of the nematode *C. elegans*. *Proc. Natl. Acad. Sci., USA* **85:** 7821-7825.

Coulson, A., Waterston, R., Kiff, J., Sulston, J. and Kohara, Y. (1988). Genome linking with yeast artificial chromosomes. *Nature* **335:** 184-6.

Crow, Y. J. and Tolmie, J. L. (1998). Recurrence risks in mental retardation. *J Med Genet* **35:** 177-82.

Daniel, J. M., Ireton, R. C. and Reynolds, A. B. (2001). Monoclonal antibodies to Kaiso: a novel transcription factor and p120ctn-binding protein. *Hybridoma* **20:** 159-66.

Dehal, P., Predki, P., Olsen, A. S., Kobayashi, A., Folta, P., Lucas, S., *et al.* (2001). Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293:** 104-11.

Delbruck, S., Sonneborn, A., Gerads, M., Grablowitz, A. H. and Ernst, J. F. (1997). Characterization and regulation of the genes encoding ribosomal proteins L39 and S7 of the human pathogen Candida albicans. *Yeast* **13:** 1199-210.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., *et al.* (1998). A physical map of 30,000 human genes. *Science* **282:** 744-6.

Deloukas, P., Matthews, L. H., Ashurst, J., Burton, J., Gilbert, J. G., Jones, M., *et al.* (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414:** 865-71.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., *et al.* (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152-4.

Dietz-Band, J. N., Turco, A. E., Willard, H. F., Vincent, A., Skolnick, M. H. and Barker, D. F. (1990). Isolation, characterization, and physical localization of 33 human X-chromosome RFLP markers. *Cytogenet Cell Genet* **54:** 137-41.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., *et al.* (1995). An integrated physical map of human chromosome 16. *Nature* **377:** 335-65.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., *et al.* (1987). A genetic linkage map of the human genome. *Cell* **51:** 319-37.

Drayna, D. and White, R. (1985). The genetic linkage map of the human X chromosome. *Science* **230:** 753-8.

Driever, W., Solnica-Krezel, L., Schier, A. F., Neuhauss, S. C., Malicki, J., Stemple, D. L., *et al.* (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123:** 37-46.

Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M., *et al.* (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res* **10:** 1304-6.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., Hunt, A. R., Collins, J. E., *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.

Duret, L. and Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **7:** 399-406.

Ekker, M., Fritz, A. and Westerfield, M. (1992). Identification of two families of satellite-like repetitive DNA sequences from the zebrafish (Brachydanio rerio). *Genomics* **13:** 1169-73.

Elgar, G., Clark, M. S., Meek, S., Smith, S., Warner, S., Edwards, Y. J., *et al.* (1999). Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning. *Genome Res* **9:** 960-71.

Eliceiri, G. L. (1999). Small nucleolar RNAs. *Cell Mol Life Sci* **56:** 22-31.

Evans, M. J., Carlton, M. B. and Russ, A. P. (1997). Gene trapping and functional genomics. *Trends Genet* **13:** 370-4.

Ewing, B. and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25:** 232-4.

Feinberg, A. P. and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* **132:** 6-13.

Fields, C., Adams, M. D., White, O. and Venter, J. C. (1994). How many genes in the human genome? *Nat Genet* **7:** 345-6.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496-512.

Foote, S., Vollrath, D., Hilton, A. and Page, D. C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* **258:** 60-6.

Frattini, A., Faranda, S., Bagnasco, L., Patrosso, C., Nulli, P., Zucchi, I., *et al.* (1997). Identification of a new member (ZNF183) of the Ring finger gene family in Xq24-25. *Gene* **192:** 291-8.

Gates, M. A., Kim, L., Egan, E. S., Cardozo, T., Sirotkin, H. I., Dougan, S. T., *et al.* (1999). A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. *Genome Res* **9:** 334-47.

Gecz, J., Barnett, S., Liu, J., Hollway, G., Donnelly, A., Eyre, H., *et al.* (1999). Characterization of the human glutamate receptor subunit 3 gene (GRIA3), a candidate for bipolar disorder and nonspecific X-linked mental retardation. *Genomics* **62:** 356-68.

Gemmill, R. M., Chumakov, I., Scott, P., Waggoner, B., Rigault, P., Cypser, J., *et al.* (1995). A second-generation YAC contig map of human chromosome 3. *Nature* **377:** 299-319.

Gerdes, D., Wehling, M., Leube, B. and Falkenstein, E. (1998). Cloning and tissue expression of two putative steroid membrane receptors. *Biol Chem* **379:** 907-11.

Gilbert, W. (1992). The Codes of Codes. Kelves, D. J. and Hood, L. Harvard University Press.

Giraud, S., Bonod-Bidaud, C., Wesolowski-Louvel, M. and Stepien, G. (1998). Expression of human ANT2 gene in highly proliferative cells: GRBOX, a new transcriptional element, is involved in the regulation of glycolytic ATP import into mitochondria. *J Mol Biol* **281:** 409-18.

Gish, W. and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Genetics* **3**: 266-72.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., *et al.* (1996). Life with 6000 genes. *Science* **274:** 546, 563-7.

Goss, S. J. and Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature* **255:** 680-4.

Gottgens, B., Barton, L. M., Gilbert, J. G., Bench, A. J., Sanchez, M. J., Bahn, S., *et al.* (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol* **18:** 181-6.

Gottgens, B., Gilbert, J. G., Barton, L. M., Grafham, D., Rogers, J., Bentley, D. R., *et al.* (2001). Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res* **11:** 87-97.

Graves, J. A. (1996). Mammals that break the rules: genetics of marsupials and monotremes. *Annu Rev Genet* **30:** 233-60.

Green, E. D. and Olson, M. V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping. *Science* **250:** 94-8.

Green, P. (1997). Against a whole-genome shotgun. *Genome Res* **7:** 410-7.

Gregg, R. G., Palmer, C., Kirkpatrick, S. and Simantel, A. (1996). Localization of a non-specific X-linked mental retardation gene, MRX23, to Xq23-q24. *Hum Mol Genet* **5:** 411-4.

Gregory, S. G., Howell, G. R. and Bentley, D. R. (1997). Genome mapping by fluorescent fingerprinting. *Genome Res* **7:** 1162-8.

Gubler, U. and Hoffman, B. J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25:** 263-9.

Guigo, R., Agarwal, P., Abril, J. F., Burset, M. and Fickett, J. W. (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10:** 1631-42.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., *et al.* (1994). The 1993-94 Genethon human genetic linkage map. *Nat Genet* **7:** 246-339.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., *et al.* (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5:** 339-46.

Haffter, P., Granato, M., Brand, M., Mullins, M. C., Hammerschmidt, M., Kane, D. A., *et al.* (1996). The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development* **123:** 1-36.

Hannenhalli, S. and Levy, S. (2001). Promoter prediction in the human genome. *Bioinformatics* **17:** S90-6.

Hardison, R., Xu, J., Jackson, J., Mansberger, J., Selifonova, O., Grotch, B., *et al.* (1993). Comparative analysis of the locus control region of the rabbit beta-like gene cluster: HS3 increases transient expression of an embryonic epsilon-globin gene. *Nucleic Acids Res* **21:** 1265-72.

Hardison, R., Slightom, J. L., Gumucio, D. L., Goodman, M., Stojanovic, N. and Miller, W. (1997). Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205:** 73-94.

Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16:** 369-72.

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405:** 311-9.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., *et al.* (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* **6:** 807-28.

Holland, J., Coffey, A. J., Giannelli, F. and Bentley, D. R. (1993). Vertical integration of cosmid and YAC resources for interval mapping on the X-chromosome. *Genomics* **15:** 297-304.

Huang, S.-H., Yang, A. Y. and Holcenberg, J. (1993). Amplification of gene ends from gene libraries by polymerase chain reaction with single-sided specficity. White, B. A. Humana Press.

Hudson, L. D., Friedrich, V. L., Jr., Behar, T., Dubois-Dalcq, M. and Lazzarini, R. A. (1989). The initial events in myelin synthesis: orientation of proteolipid protein in the plasma membrane of cultured oligodendrocytes. *J Cell Biol* **109:** 717-27.

Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., *et al.* (1995). An STS-based map of the human genome. *Science* **270:** 1945-54.

Hudson, T. J., Church, D. M., Greenaway, S., Nguyen, H., Cook, A., Steen, R. G., *et al.* (2001). A radiation hybrid map of mouse genes. *Nat Genet* **29:** 201-5.

IHGSC (International Human Genome Sequencing Consortium). (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Imai, T. and Olson, M. V. (1990). Second-generation approach to the construction of yeast artificial-chromosome libraries. *Genomics* **8:** 297-303.

Inoue, K., Osaka, H., Imaizumi, K., Nezu, A., Takanashi, J., Arii, J., *et al.* (1999). Proteolipid protein gene duplications causing Pelizaeus-Merzbacher disease: molecular mechanism and phenotypic manifestations. *Ann Neurol* **45:** 624-32.

Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., *et al.* (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6:** 84-9.

Izsvak, Z., Ivics, Z. and Hackett, P. B. (1995). Characterization of a Tc1-like transposable element in zebrafish (Danio rerio). *Mol Gen Genet* **247:** 312-22.

Izsvak, Z., Ivics, Z., Garcia-Estefania, D., Fahrenkrug, S. C. and Hackett, P. B. (1996). DANA elements: a family of composite, tRNA-derived short interspersed DNA elements associated with mutational activities in zebrafish (Danio rerio). *Proc Natl Acad Sci U S A* **93:** 1077-81.

Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H. and Hackett, P. B. (1999). Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J Mol Evol* **48:** 13-21.

Jackson, M. S. and Strachan, T. (1996). Human Genome Evolution. Oxford Bios.

Jang, W., Hua, A., Spilson, S. V., Miller, W., Roe, B. A. and Meisler, M. H. (1999). Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res* **9:** 53-61.

Jeppesen, P. and Turner, B. M. (1993). The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* **74:** 281-9.

Justice, M. J. (2000). Capitalizing on large-scale mouse mutagenesis screens. *Nat Rev Genet* **1:** 109-15.

Kannan, K., Stewart, R. M., Bounds, W., Carlsson, S. R., Fukuda, M., Betzing, K. W., *et al.* (1996). Lysosome-associated membrane proteins h-LAMP1 (CD107a) and h-LAMP2 (CD107b) are activation-dependent cell surface glycoproteins in human peripheral blood mononuclear cells which mediate cell adhesion to vascular endothelium. *Cell Immunol* **171:** 10-9.

Kaplan, J. C., Kahn, A. and Chelly, J. (1992). Illegitimate transcription: its use in the study of inherited disease. *Hum Mutat* **1:** 357-60.

Kendall, E., Evans, W., Jin, H., Holland, J. and Vetrie, D. (1997). A complete YAC contig and cosmid interval map covering the entirety of human Xq21.33 to Xq22.3 from DXS3 to DXS287. *Genomics* **43:** 171-82.

Khan, A. S., Wilcox, A. S., Hopkins, J. A. and Sikela, J. M. (1991). Efficient double stranded sequencing of cDNA clones containing long poly(A) tails using anchored poly(dT) primers. *Nucleic Acids Res* **19:** 1715.

Khan, A. S., Wilcox, A. S., Polymeropoulos, M. H., Hopkins, J. A., Stevens, T. J., Robinson, M., *et al.* (1992). Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nat Genet* **2:** 180-5.

Kim, U. J., Shizuya, H., de Jong, P. J., Birren, B. and Simon, M. I. (1992). Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20:** 1083-5.

Kinoshita, M., Kumar, S., Mizoguchi, A., Ide, C., Kinoshita, A., Haraguchi, T., *et al.* (1997). Nedd5, a mammalian septin, is a novel cytoskeletal component interacting with actin-based structures. *Genes Dev* **11:** 1535-47.

Kohara, Y., Akiyama, K. and Isono, K. (1987). The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50:** 495-508.

Koken, M. H., Hoogerbrugge, J. W., Jasper-Dekker, I., de Wit, J., Willemsen, R., Roest, H. P., *et al.* (1996). Expression of the ubiquitin-conjugating DNA repair enzymes HHR6A and B suggests a role in spermatogenesis and chromatin modification. *Dev Biol* **173:** 119-32.

Koller, B. H. and Smithies, O. (1992). Altering genes in animals by gene targeting. *Annu Rev Immunol* **10:** 705-30.

Koop, B. F. and Hood, L. (1994). Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat Genet* **7:** 48-53.

Kozak, M. (1991). Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* **266:** 19867-70.

Krauter, K., Montgomery, K., Yoon, S. J., LeBlanc-Straceski, J., Renault, B., Marondel, I., *et al.* (1995). A second-generation YAC contig map of human chromosome 12. *Nature* **377:** 321-33.

Lapenta, V., Chiurazzi, P., van der Spek, P., Pizzuti, A., Hanaoka, F. and Brahe, C. (1997). SMT3A, a human homologue of the S. cerevisiae SMT3 gene, maps to chromosome 21qter and defines a novel gene family. *Genomics* **40:** 362-6.

Larin, Z., Monaco, A. P. and Lehrach, H. (1991). Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc Natl Acad Sci U S A* **88:** 4123-7.

Lehrke, R. G. (1974). X-linked mental retardation and verbal disability. *Birth Defects Orig Artic Ser* **10:** 1-100.

Levine, A. and Durbin, R. (2001). A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res* **29:** 4006-13.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25:** 239-40.

Lin, C. S., Park, T., Chen, Z. P. and Leavitt, J. (1993). Human plastin genes. Comparative gene structure, chromosome location, and differential expression in normal and neoplastic cells. *J Biol Chem* **268:** 2781-92.

Little, P. F., Flavell, R. A., Kooter, J. M., Annison, G. and Williamson, R. (1979). Structure of the human fetal globin gene locus. *Nature* **278:** 227-31.

Luciakova, K., Hodny, Z., Barath, P. and Nelson, B. D. (2000). In vivo mapping of the human adenine nucleotide translocator-2 (ANT2) promoter provides support for regulation by a pair of proximal Sp1-activating sites and an upstream silencer element. *Biochem J* **352:** 519-23.

Lyon, M. F. (1961). Gene action in the X chromosome of the mouse. *Nature* **190:** 372-373.

Lyon, M. F. (1998). X-chromosome inactivation: a repeat hypothesis. *Cytogenet Cell Genet* **80:** 133-7.

Mackey, D. A., Buttery, R. G., Wise, G. M. and Denton, M. J. (1991). Description of X-linked megalocornea with identification of the gene locus. *Arch Ophthalmol* **109:** 829-33.

Maestrini, E., Rivella, S., Tribioli, C., Purtilo, D., Rocchi, M., Archidiacono, N., *et al.* (1990). Probes for CpG islands on the distal long arm of the human X chromosome are clustered in Xq24 and Xq28. *Genomics* **8:** 664-70.

Makalowski, W., Zhang, J. and Boguski, M. S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res* **6:** 846-57.

Manoni, M., Tribioli, C., Lazzari, B., DeBellis, G., Patrosso, C., Pergolizzi, R., *et al.* (1991). The nucleotide sequence of a CpG island demonstrates the presence of the first exon

of the gene encoding the human lysosomal membrane protein lamp2 and assigns the gene to Xq24. *Genomics* **9:** 551-4.

Mathias, N., Johnson, S. L., Winey, M., Adams, A. E., Goetsch, L., Pringle, J. R., *et al.* (1996). Cdc53p acts in concert with Cdc4p and Cdc34p to control the G1-to-S-phase transition and identifies a conserved family of proteins. *Mol Cell Biol* **16:** 6634-43.

McKusick, V. A. (1998). Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders. John Hopkins University Press.

Mei, X., Singh, I. S., Erlichman, J. and Orr, G. A. (1997). Cloning and characterization of a testis-specific, developmentally regulated A-kinase-anchoring protein (TAKAP-80) present on the fibrous sheath of rat sperm. *Eur J Biochem* **246:** 425-32.

Metscher, B. D. and Ahlberg, P. E. (1999). Zebrafish in context: uses of a laboratory model in comparative studies. *Dev Biol* **210:** 1-14.

Miller, W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17:** 391-7.

Mills, A. A., Zheng, B., Wang, X. J., Vogel, H., Roop, D. R. and Bradley, A. (1999). p63 is a p53 homologue required for limb and epidermal morphogenesis. *Nature* **398:** 708-13.

Monaco, A. P. and Larin, Z. (1994). YACs, BACs, PACs and MACs: artificial chromosomes as research tools. *Trends Biotechnol* **12:** 280-6.

Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., *et al.* (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* **265:** 2049-54.

Nagaraja, R., MacMillan, S., Kere, J., Jones, C., Griffin, S., Schmatz, M., *et al.* (1997). X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res* **7:** 210-22.

Nagaraja, R., MacMillan, S., Jones, C., Masisi, M., Pengue, G., Porta, G., *et al.* (1998). Integrated YAC/STS physical and genetic map of 22.5 Mb of human Xq24-q26 at 56-kb inter-STS resolution. *Genomics* **52:** 247-66.

Nakamura, Y., Julier, C., Wolff, R., Holm, T., O'Connell, P., Leppert, M., *et al.* (1987). Characterization of a human 'midisatellite' sequence. *Nucleic Acids Res* **15:** 2537-47.

Naylor, J. A., Buck, D., Green, P., Williamson, H., Bentley, D. and Giannelli, F. (1995). Investigation of the factor VIII intron 22 repeated region (int22h) and the associated inversion junctions. *Hum Mol Genet* **4:** 1217-24.

Neri, G. and Chiurazzi, P. (1999). X-linked mental retardation. *Adv Genet* **41:** 55-94.

Nizetic, D., Gellen, L., Hamvas, R. M., Mott, R., Grigoriev, A., Vatcheva, R., *et al.* (1994). An integrated YAC-overlap and 'cosmid-pocket' map of the human chromosome 21. *Hum Mol Genet* **3:** 759-70.

Nourbakhsh, M., Oumard, A., Schwarzer, M. and Hauser, H. (2000). NRF, a nuclear inhibitor of NF-kappaB proteins silencing interferon-beta promoter. *Eur Cytokine Netw* **11:** 500-1.

O'Brien, S. J., Eisenberg, J. F., Miyamoto, M., Hedges, S. B., Kumar, S., Wilson, D. E., *et al.* (1999). Genome maps 10. Comparative genomics. Mammalian radiations. Wall chart. *Science* **286:** 463-78.

Ohno, S. (1967). Sex chromosomes and sex linked genes. Springer.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., *et al.* (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci U S A* **83:** 7826-30.

Osaka, F., Kawasaki, H., Aida, N., Saeki, M., Chiba, T., Kawashima, S., *et al.* (1998). A new NEDD8-ligating system for cullin-4A. *Genes Dev* **12:** 2263-8.

Panning, B. and Jaenisch, R. (1998). RNA and the epigenetic regulation of X chromosome inactivation. *Cell* **93:** 305-8.

Papadakis, M. N. and Patrinos, G. P. (1999). Contribution of gene conversion in the evolution of the human beta-like globin gene family. *Hum Genet* **104:** 117-25.

Pestov, N. B., Adams, G., Shakhparonov, M. I. and Modyanov, N. N. (1999). Identification of a novel gene of the X,K-ATPase beta-subunit family that is predominantly expressed in skeletal and heart muscles. *FEBS Lett* **456:** 243-8.

Postlethwait, J. H., Yan, Y. L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., *et al.* (1998). Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* **18:** 345-9.

Rabbitts, T. H. (1976). Bacterial cloning of plasmids carrying copies of rabbit globin messenger RNA. *Nature* **260:** 221-5.

Rastan, S. (1983). Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre. *J Embryol Exp Morphol* **78:** 1-22.

Rastan, S. (1994). X chromosome inactivation and the Xist gene. *Curr Opin Genet Dev* **4:** 292-7.

Richards, R. I., Holman, K., Kozman, H., Kremer, E., Lynch, M., Pritchard, M., *et al.* (1991). Fragile X syndrome: genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. *J Med Genet* **28:** 818-23.

Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., *et al.* (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res* **18:** 2887-90.

Roberts, R. G., Barby, T. F., Manners, E., Bobrow, M. and Bentley, D. R. (1991). Direct detection of dystrophin gene rearrangements by analysis of dystrophin mRNA in peripheral blood lymphocytes. *Am J Hum Genet* **49:** 298-310.

Roest Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., *et al.* (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet* **25:** 235-8.

Ross, M. and Langford, C. F. (1997). The use of flow-sorted chromosomes in genome mapping. In *Genome Mapping: a practical approach.* Dear, P. H. IRL Press at Oxford University Press. 165-184.

Rubin, E. M. (2001). Comparing Species. *Nature* **409:** 820-821.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., *et al.* (2000). Comparative genomics of the eukaryotes. *Science* **287:** 2204-15.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). Molecular cloning; a laboratory manual. Cold Spring Harbor Laboratory Press.

Sanchez, R. and Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* **Suppl:** 50-8.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., *et al.* (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265:** 687-95.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-7.

Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., *et al.* (1978). The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* **125:** 225-46.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. and Petersen, G. B. (1982). Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162:** 729-73.

Scherf, M., Klingenhoff, A. and Werner, T. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* **297:** 599-606.

Schiebel, K., Mertz, A., Winkelmann, M., Nagaraja, R. and Rappold, G. (1994). Localization of the adenine nucleotide translocase gene ANT2 to chromosome Xq24-q25 with tight linkage to DXS425. *Genomics* **24:** 605-6.

Schnedl, W., Mikelsaar, A. V., Breitenbach, M. and Dann, O. (1977). DIPI and DAPI: fluorescence banding with only negliglible fading. *Hum Genet* **36:** 167-72.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., *et al.* (1996). A gene map of the human genome. *Science* **274:** 540-6.

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., *et al.* (2000). PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* **10:** 577-86.

Shashidharan, P., Michaelidis, T. M., Robakis, N. K., Kresovali, A., Papamatheakis, J. and Plaitakis, A. (1994). Novel human glutamate dehydrogenase expressed in neural and testicular tissues and encoded by an X-linked intronless gene. *J Biol Chem* **269:** 16971-6.

Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y., *et al.* (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89:** 8794-7.

Simmons, D. L. 1993.Cloning cell surface molecules by transient expression in mammalian cells. IRL Press at Oxford Univeristy PressOxford. 93-127

Sloan, J. L. and Mager, S. (1999). Cloning and functional expression of a human Na(+) and Cl(-)-dependent neutral and cationic amino acid transporter B(0+). *J Biol Chem* **274:** 23740-5.

Solovyev, V. V., Salamov, A. A. and Lawrence, C. B. (1995). Identification of human gene structure using linear discriminant functions and dynamic programming. *Proc Int Conf Intell Syst Mol Biol* **3:** 367-75.

Sonnhammer, E. L. and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1-10.

Srivastava, A. K., McMillan, S., Jermak, C., Shomaker, M., Copeland-Yates, S. A., Sossey-Alaoui, K., *et al.* (1999). Integrated STS/YAC physical, genetic, and transcript map of human Xq21.3 to q23/q24 (DXS1203-DXS1059). *Genomics* **58:** 188-201.

Staden, R. (1984). Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Res* **12:** 521-38.

Steingruber, H. E., Dunham, A., Coffey, A. J., Clegg, S. M., Howell, G. R., Maslen, G. L., *et al.* (1999). High-resolution landmark framework for the sequence-ready mapping of Xq23-q26.1. *Genome Res* **9:** 751-62.

Stewart, E. A., McKusick, K. B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A., *et al.* (1997). An STS-based radiation hybrid map of the human genome. *Genome Res* **7:** 422-33.

Takagi, N. (1974). Differentiation of X chromosomes in early female mouse embryos. *Exp Cell Res* **86:** 127-35.

Thanaraj, T. A. and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res* **29:** 2581-93.

Toniolo, D. and D'Adamo, P. (2000). X-linked non-specific mental retardation. *Curr Opin Genet Dev* **10:** 280-5.

Touchman, J. W., Dehejia, A., Chiba-Falek, O., Cabin, D. E., Schwartz, J. R., Orrison, B. M., *et al.* (2001). Human and mouse alpha-synuclein genes: comparative genomic sequence analysis and identification of a novel gene regulatory element. *Genome Res* **11:** 78-86.

Trofatter, J. A., Dlouhy, S. R., DeMyer, W., Conneally, P. M. and Hodes, M. E. (1989). Pelizaeus-Merzbacher disease: tight linkage to proteolipid protein gene exon variant. *Proc Natl Acad Sci U S A* **86:** 9427-30.

Trofatter, J. A., Long, K. R., Murrell, J. R., Stotler, C. J., Gusella, J. F. and Buckler, A. J. (1995). An expression-independent catalog of genes from human chromosome 22. *Genome Res* **5:** 214-24.

Tsukada, S., Saffran, D. C., Rawlings, D. J., Parolini, O., Allen, R. C., Klisak, I., *et al.* (1993). Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. *Cell* **72:** 279-90.

Tyson, J., Bellman, S., Newton, V., Simpson, P., Malcolm, S., Pembrey, M. E., *et al.* (1996). Mapping of DFN2 to Xq22. *Hum Mol Genet* **5:** 2055-60.

Uberbacher, E. C. and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci U S A* **88:** 11261-5.

Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science* **280:** 1540-2.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304-51.

Vetrie, D. (1993). Isolation of the defective gene in X linked agammaglobulinaemia. *J Med Genet* **30:** 452-3.

Vetrie, D., Kendall, E., Coffey, A., Hassock, S., Collins, J., Todd, C., *et al.* (1994). A 6.5-Mb yeast artificial chromosome contig incorporating 33 DNA markers on the human X chromosome at Xq22. *Genomics* **19:** 42-7.

Vogel, A. M. and Gerster, T. (1999). Promoter activity of the zebrafish bhikhari retroelement requires an intact activin signaling pathway. *Mech Dev* **85:** 133-46.

Walhout, A. J., Boulton, S. J. and Vidal, M. (2000). Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* **17:** 88-94.

Way, M., Sanders, M., Chafel, M., Tu, Y. H., Knight, A. and Matsudaira, P. (1995). beta-Scruin, a homologue of the actin crosslinking protein scruin, is localized to the acrosomal vesicle of Limulus sperm. *J Cell Sci* **108:** 3155-62.

Weber, J. L. and Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Res* **7:** 401-9.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., *et al.* (1992). A second-generation linkage map of the human genome. *Nature* **359:** 794-801.

Willard, H. F. (1996). X chromosome inactivation, XIST, and pursuit of the X-inactivation center. *Cell* **86:** 5-7.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., *et al.* (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28:** 316-9.

Wyman, A. R. and White, R. (1980). A highly polymorphic locus in human DNA. *Proc Natl Acad Sci U S A* **77:** 6754-8.

Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci U S A* **94:** 565-8.

Zhang, P., Li, M. Z. and Elledge, S. J. (2002). Towards genetic genome projects: genomic library screening and gene-targeting vector construction in a single step. *Nat Genet* **30:** 31-9.

Zollman, S., Godt, D., Prive, G. G., Couderc, J. L. and Laski, F. A. (1994). The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in Drosophila. *Proc Natl Acad Sci U S A* **91:** 10717-21.